University of Alberta

Modeling L1/L2 interactions in the perception and production of English vowels by Mandarin L1 speakers: A training study

by

Ronald Irvin Thomson ©

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Linguistics

Edmonton, Alberta
Fall 2007

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-33079-1
Our file Notre référence
ISBN: 978-0-494-33079-1

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

I close my eyes
And think of all the ways you feed my mind
with all that's true
to make it easy to remember
And when I do
the more I long for times when I'm alone in you
And know I'll die
if I don't keep your word forever

Something I ever needed, appreciated
And now I've made it
No longer tainted
Because of you

- From "Because of You", Spyglass Blue (2001).

For my wife Marcela

Abstract

This dissertation proposes a new statistical pattern recognition model for measuring crosslinguistic vowel similarity and applies it to measuring differences between English and Mandarin vowels. The model incorporates multidimensional acoustic information previously found to be important for vowel identification. Previous research has trained similar statistical models on vowel production values from a single language and used these trained models to determine how new production tokens from a second language might be classified in terms of the first language. The current statistical model was trained on production values from both languages being compared, Mandarin and English. New production tokens from both languages were then tested against this statistical model to determine the extent to which production tokens in one language were misidentified as members of opposing language categories. The degree of misidentification across languages provides a metric for determining how similar Mandarin and English vowels are to each other. From this, explicit predictions were made concerning how Mandarin speakers would identify and produce L2 English vowels. In a training study, twenty-six L2 learners were trained under three conditions to identify ten English vowels. Changes in the learners' ability to identify and produce these English vowels were used to measure the effect of instruction as well as predictions of the statistical model of Mandarin-English vowel similarity. Results indicate that this statistical approach to measuring crosslinguistic similarity can be used to quite accurately predict the behaviour of adult Mandarin learners of L2 English. Furthermore, the results demonstrate that under certain training conditions, Mandarin learners of English are able to improve in their ability to identify and produce English vowels.

Acknowledgments

I am thankful to a great number of people for their encouragement, support and friendship throughout the years that culminated in this dissertation. First and foremost, I am indebted to my co-supervisors Terrance Nearey and Tracey Derwing. Terry's expertise and guidance has been invaluable. He has helped me to complete a program of study that I could not have done elsewhere. I am grateful for the interest he took in what I wanted to pursue and the many, many hours he has spent making my research better. His willingness to share his acoustic analysis software and tweak it to meet my needs has literally saved me hundreds of hours in measuring and analyzing thousands of vowel recordings. I am equally grateful to Tracey, who was the first to encourage me to pursue a PhD, after completing an M.Ed. under her tutelage. Her generosity throughout my years of work in the TESL Research Centre is immeasurable. Working in Tracey's lab allowed me to hone my skills as a researcher – something no amount of classroom training could have accomplished.

I am also very thankful for the contributions made by my other committee members: Susan Guion, Karen Pollock and Robert Kirchner. Thank you all for the thoughtful comments and the great amount of time and effort you put into helping me make my research stronger. Thanks to Robert for being a source of encouragement throughout my doctoral studies.

I owe a huge debt of gratitude to the anonymous participants in my study, both at the University of Alberta and at NorQuest College. The Mandarin learners of English in particular trusted me enough to volunteer for many hours of training. I am grateful to the administration of NorQuest College, especially Anna DeLuca and Dorte Weber, for allowing me to conduct my research on their campus and to the teachers who helped me to recruit students. I also need to give special thanks to Milorad Zivkovic for permitting me to access the computer facilities and making sure my needs were being met.

Other academic types who have provided encouragment and influenced me along the way include Leila Ranta, Sally Rice, Marian Rossiter, Bill Dunn, Johanne Paradis, Kim Noels, Gary Libben and John Hogan at the U of A, and Murray Munro from Simon Fraser University. Thanks are due to Debra Elliot and Jana Tomasovic in the Department of Linguistics for their administrative support and friendly smiles.

Fellow students who have made an impact include Anita Leander, Philip Hallstrom, Becca Avery, Holly Leslie, Bernie Hendricks, Lesley Dudley, Jennifer Foote, Talia Isaacs, Linda Manimtim, Nicole Ringuette, Peter Myhre and Geoff Morrison.

In addition, I must also acknowledge my instructors and classmates from the first summer I spent at the Summer Institute of Linguistics (SIL) at the University of North Dakota. I cannot imagine a more supportive, yet intense context for beginning my formal study of linguistics. I am particularly thankful for my hellish experience in articulatory phonetics, which caused me to ask many questions, eventually leading me to acoustic phonetics where I found my intellectual salvation.

I am thankful to my family as well. My initial interest in second language acquisition was the result of being born to a dad who left his linguistics books in every nook and cranny of our many homes. Some of my earliest memories are of his attempts at learning Blackfoot and later, Urdu. Even today he continues to live his life as his own participant in a seemingly endless linguistic experiment. There is no doubt that his life's work and my exposure to it has been a major influence on my own intellectual formation. Thanks dad. Thanks mom for putting up with two linguists in the family and for your general encouragement. Thanks to Ambrose for being a great roommate during my first few years in Edmonton, and to Holly for reminding me from time to time that I do still have a biological family. I am also thankful for the strong support offered from a distance by my grandparents. Thanks for being proud of me, despite not really knowing what I'm doing. Special thanks are due to Marcela, my wife. Thank you for bearing so much! Your support during the past year has been incredible. Thank you for being so patient and understanding of the many long hours I've had to spend away from home. I know that you are probably happier than I am that this dissertation has come to an end. I look forward to truly beginning our life together.

My research was made possible through financial support from a Province of Alberta Graduate Fellowship, a Killam Doctoral Fellowship and a SSHRC Doctoral Fellowship, as well as SSHRC research grants held by both of my co-supervisors.

Finally, thank you Jesus for making me who I am and giving me the ability and perseverance to investigate the wonder that is language. Any glory remains yours.

Chapter 1. Second language speech learning 1

1.1. Fundamental differences in L1 and L2 speech learning 2
  1.1.1. Critical Period Hypothesis 2
  1.1.2. L1 speech learning 4
  1.1.3. L2 speech learning 7
1.2. The role of learners' L1 in L2 speech learning 8
1.3. Current models of L2 speech perception 15
1.4. The relationship between perception and production in L2 speech learning ... 19
1.5. Attention and instructional intervention in L2 speech learning 21
1.6. Summary 28

Chapter 2. Measuring crosslinguistic vowel similarity 29

2.1. A crosslinguistic similarity continuum 29
2.2. Past approaches to crosslinguistic similarity 33
2.3. A statistical pattern recognition approach to L2 phonological learning 38
  2.3.1. Measuring crosslinguistic similarity 38
  2.3.2. The Metamodel and L2 phonological representations 44
2.4. Summary 45

Chapter 3. A comparison of English and Mandarin vowel systems 46

3.1. Method 49
  3.1.1. Mandarin and English vowel inventory selection 49
  3.1.2. Speakers 53
  3.1.3. Procedure 53
  3.1.4. Data analysis 56
3.2. Results 61
3.3. Discussion 68
  3.3.1. Mandarin and English models 68
  3.3.2. Metamodel 70
  3.3.3. English x Mandarin and Mandarin x English models 70
  3.3.4. Predictions for Mandarin L1 learners of English 71

Chapter 4. Training and its effect on perception and production 75

4.1. Research Questions 77
4.2. Method 77
  4.2.1. Participants 77
  4.2.2. Stimuli 79
  4.2.3. Procedure 82
  4.2.4. Data analysis 87
4.3. Results 88
  4.3.1. Identification tests 88
  4.3.2. Vowel production test 110
  4.3.3. Relationship between identification and production results 117
4.4. Discussion 131
  4.4.1. Perception 131
  4.4.2. Production 134

Chapter 5. Testing predictions of L2 speech learning models 137

5.1. Predictions 138
  5.1.1. Identification confusion patterns 138
  5.1.2. Production confusion patterns 141
5.2. Method 143
  5.2.1. L2 English identification data 143
  5.2.2. L2 English production data 143
5.3. Results 143
  5.3.1. Identification confusion patterns 143
  5.3.2. Production confusion patterns 150
5.4. Discussion 157
  5.4.1. Identification Confusion Patterns 157
  5.4.2. Production Confusion Patterns 159
  5.4.3. Relationship between perception and production 161

Chapter 6. A brief comparison of training with naturalistic L2 vowel development 164

6.1. Research Questions 164

6.2. Method........165
6.2.1. Speakers........165
6.2.2. Procedure........165
6.2.3. Data analysis........166
6.3. Results........167
6.3.1. Naturalistic L2 English learners' vowel production data........167
6.3.2. Naturalistic vowel learning versus instructed vowel learning........169
6.4. Discussion........174

Chapter 7. General summary and discussion........177

7.1. Measuring crosslinguistic vowel similarity using the Metamodel........177
7.2. The effect of training on L2 vowel perception and production........181
7.2.1. Comparing differences in vowel identification training conditions..181
7.2.2. Transfer of training to new phonetic contexts........185
7.2.3. Transfer of training to novel speaker's voice........185
7.2.4. Transfer of vowel identification training to production........186
7.3. The effect of the L1 on L2 speech learning........187
7.4. Implications for L2 speech learning models........189
7.5. Pedagogical implications........191
7.6. Further research........193

Bibliography........195
Appendix 1. Alternate pattern recognition model results: Mandarin CV with vowel duration; English CV without vowel duration; Metamodel CV with vowel duration........208
Appendix 2. Details of training study participants........212
Appendix 3. Details of individual CV production tokens used for Generalization and Production test stimuli........214
Appendix 4. Presentation order of English vowel categories in the training Demo Mode........226
Appendix 5. Detailed statistics for the L2 English vowel training study........227

Appendix 6. Detailed statistics for the L2 English production results when vowel duration is included a variable in the CV English model 240
Appendix 7. Identification confusion matrixes for Natural and Lengthened Vowel tests at Time 1 and Time 2 by each subset of participants who took each test at both times. 246
Appendix 8. Detailed statistics for the naturalistic L2 English vowel production data.248
Appendix 9. Alternate statistics for the naturalistic L2 English vowel data and comparison with the training study data when vowel duration is excluded as a variable...253

List of Tables

Table 3.1. Syllables used for elicited imitation of seven target Mandarin vowels. 54
Table 3.2. Syllables used for elicited imitation of ten target English vowels..................55
Table 3.3. Mandarin Model trained and tested on L1 Mandarin productions. 62
Table 3.4. English Model trained and tested on L1 English productions with................63
Table 3.5. Metamodel trained and tested on L1 English and L1 Mandarin productions..64
Table 3.6. Percentage of English vowel tokens (n = 40 per vowel) with APPs of >.05 of being classified as a Mandarin vowel in the Metamodel. 65
Table 3.7. Percentage of Mandarin vowel tokens with APPs of >.05 of being classified as an English vowel in the Metamodel. 66
Table 3.8. English items tested on the Mandarin Model. 67
Table 3.9. Mandarin items tested on the English Model 68
Table 4.1. Order of training and testing phases with approximate timeline in parentheses. 83
Table 4.2. Mean % correct vowel identification scores and standard deviations on Generalization Test by CV context, Stimulus Voice, Training Group and Time. 89
Table 4.3. Mean % correct vowel identification scores and standard deviations by Lengthened Vowel vs Natural Vowel test , Training Group and Time. 96
Table 4.4. Mean % correct identification scores and standard deviations pooled across all subjects for each subclass of tokens. 101
Table 4.5. Mean % correct vowel identification scores and standard deviations on the Generalization test by CV context, and stimulus Voice at Time 1, Time 2 and Delayed post-test. 102
Table 4.6. Mean % correct identification scores and standard deviations on the Natural Vowel test by Training Group and Time 108
Table 4.7. Mean % correct vowel production recognition scores and standard deviations on the production test as recognized by the English Model............................... 111
Table 4.8. Mean vowel duration and standard deviation for productions of /u/ and /ʊ/, in terms of how the intended vowel was recognized by the English Model........... 125

Table 4.9. Mean vowel duration and standard deviation for productions of /ɒ/ and /ʌ/, in terms of how the intended vowel was recognized by the English Model........... 129

Table 4.10. Vowel duration for each stimulus CV, by stimulus voice........................... 130

Table 5.1. Predictions regarding L2 English vowel identification patterns based on the Metamodel analysis of Mandarin and English vowels in Chapter 3. 140

Table 5.2. Predictions regarding L2 English vowel production patterns based on the Metamodel analysis of Mandarin and English vowels in Chapter 3. 142

Table 5.3. Listener identification of English /b, pV/ stimuli on the Natural Vowel Test at Time 1 and Time 2.. 144

Table 5.4. Listener identification of English /b, pV/ stimuli on the Lengthened Vowel test at Time 1 and Time 2.. 145

Table 5.5. Listener identification of English /b, pV/ stimuli on Generalization Test Time 1 and Time 2, Voice 1.. 148

Table 5.6. Listener identification of English /b, pV/ stimuli on Generalization Test Time 1 and Time 2, Voice 2.. 149

Table 5.7. L2 production data tested on Metamodel excluding vowel duration as variable in response to Voice 1 stimuli. 151

Table 5.8. L2 production data tested on Metamodel excluding vowel duration as variable in response to Voice 2 stimuli. 152

Table 5.9. Percentage of L1 English productions recognized as the intended English category by the Metamodel compared with L2 productions classified by the Metamodel as the intended English vowel by voice and time. 154

Table 5.10. Mean Mahalanobis Distance scores for NS English and L2 English production tokens that were correctly identified by the English Model. 156

Table 6.1. Training and testing stimuli for ten English vowels. 166

Table 6.2. Recognition of English CVC production tokens by vowel tested against the CVC English Model trained and tested on native speaker English productions with vowel duration included as a variable... 168

Table 6.3. Mean percent correctly recognized vowel productions over time for the naturalistic group's L2 productions (top panel), in contrast to the trained group's L2 productions described in Chapters 4 and 5 (bottom panel)........................... 170

Table 6.4. Summary of mean percent correctly recognized L2 English productions for each English vowel, contrasting naturalistic vowel learning results (from Times 1-6) with trained vowel learning results (from Times 1-2) 172

Table A1.1. Mandarin Model trained and tested on native speaker Mandarin productions 208

Table A1.2. English Model trained and tested on native speaker English productions . 208

Table A1.3. Metamodel trained and tested on native speaker English and Mandarin ... 209

Table A1.4. English items tested on the Mandarin Model 210

Table A1.5. Mandarin items tested on the English Model 211

Table A2.1. Details of all English vowel training participants who completed the training portion of the study (n = 26) 212

Table A2.2. Details of all English vowel training participants who completed the delayed post-test portion of the study (n = 18) 213

Table A3.1. English Model recognition of Voice 1 /b, pV/ stimuli (vowel duration included) 214

Table A3.2. English Model recognition of Voice 2 /b, pV/ stimuli (vowel duration included 215

Table A3.3. English Model recognition of Voice 1 /g, kV/ stimuli (vowel duration included) 216

Table A3.4. English Model recognition of Voice 1 /z, sV/ stimuli (vowel duration included) 217

Table A3.5. English Model recognition of Voice 1 /b, pV/ stimuli (vowel duration excluded) 218

Table A3.6. English Model recognition of Voice 2 /b, pV/ stimuli (vowel duration excluded) 219

Table A3.7. English Model recognition of Voice 1 /g, kV/ stimuli (vowel duration excluded) 220

Table A3.8. English Model recognition of Voice 1 /z, sV/ stimuli (vowel duration excluded) 221

Table A3.9. Metamodel recognition of Voice 1 /b, pV/ stimuli (vowel duration excluded) 222

Table A3.10. Metamodel recognition of Voice 2 /b, pV/ stimuli (vowel duration excluded) 223

Table A3.11. Metamodel recognition of Voice 1 /g, kV/ stimuli (vowel duration excluded) 224

Table A3.12. Metamodel recognition of Voice 1 /z, sV/ stimuli (vowel duration excluded) 225

Table A4. Presentation order of English vowel categories in Demo Mode 226

Table A5.1. ANOVA and Multivariate Tests comparing results on Generalization Test by Time, Consonant and Vowel, for Voice 1 only. 227

Table A5.2. ANOVA and Multivariate Tests comparing results on Generalization Test by Time, Voice and Vowel, for /b, pV/ context only. 228

Table A5.3. ANOVA results on Training Vowel Identification Tests (Lengthened or Natural) by Time and Vowel 229

Table A5.4. ANOVA and Multivariate Tests comparing results on Lengthened and Natural Identification Tests by Vowel Stimulus Length and Vowel.................. 230

Table A5.5. ANOVA and Multivariate Tests comparing results on Generalization Identification Test at Time 2 with results on Generalization Delayed post-test by Consonant and Vowel for Voice 1 only. 231

Table A5.6. ANOVA and Multivariate Tests comparing results on Generalization Test at Time 2 with results on Generalization Delayed post-test by Voice and Vowel.. 232

Table A5.7. ANOVA results on Natural Vowel Identification on Test 2 and Delayed Post-test by Vowel. 233

Table A5.8. ANOVA and Multivariate Tests comparing production results by Time, Consonant pair, and Vowel, in response to Voice 1 only, tested on the English Model with vowel duration excluded as a variable. 234

Table A5.9. ANOVA and Multivariate Tests comparing production results for /b, pV/ only by Time, and Vowel, in response to Voice 1 only, tested on the English Model with vowel duration excluded as a variable. 235

Table A5.10. ANOVA and Multivariate Tests comparing production results for /g, kV/ only by Time, and Vowel, in response to Voice 1 only, tested on the English Model with vowel duration excluded as a variable. 236

Table A5.11. ANOVA and Multivariate Tests comparing production results for /z, sV/ only by Time, and Vowel, in response to Voice 1 only, tested on the English Model with vowel duration excluded as a variable. 237

Table A5.12. ANOVA and Multivariate Tests comparing production results by Time, Voice, and Vowel, for the /b, pV/ context only, tested on the English Model with vowel duration excluded as a variable. 238

Table A6.1. Mean % correct vowel production recognition scores by CV context, stimulus Voice and Time..240

Table A6.2. ANOVA and Multivariate Tests comparing production results by Time, Consonant pair, and Vowel, in response to Voice 1 only, tested on the English Model with vowel duration included as a variable. 241

Table A6.3. ANOVA and Multivariate Tests comparing production results for /b, pV/ only by Time, and Vowel, in response to Voice 1 only, tested on the English Model with vowel duration included as a variable. 242

Table A6.4. ANOVA and Multivariate Tests comparing production results for /g, kV/ only by Time, and Vowel, in response to Voice 1 only, tested on the English Model with vowel duration included as a variable. 243

Table A6.5. ANOVA and Multivariate Tests comparing production results for /z, sV/ only by Time, and Vowel, in response to Voice 1 only, tested on the English Model with vowel duration included as a variable. 244

Table A6.6. ANOVA and Multivariate Tests comparing production results by Time, Voice, and Vowel, for the /b, pV/ context only, tested on the English Model with vowel duration included as a variable..245

Table A7.1. Identification of English /b, pV/ stimuli on the Natural Vowel Test at Time 1 and Time 2, for the SVT and DVT groups only. 246

Table A7.2. Identification of English /b, pV/ stimuli on the Lengthened Vowel Test at Time 1 and Time 2, for the LVT group only..247

Table A8.1. ANOVA and Multivariate test results for naturalistic L2 English vowel production data including vowel duration as a variable. 248

Table A8.2. Naturalistic group's Time 1 L2 English CVC productions, tested on the CVC English Model with vowel duration included as a variable. 249

Table A8.3. Naturalistic group's Time 2 L2 English CVC productions, tested on the CVC English Model with vowel duration included as a variable. 249

Table A8.4. Naturalistic group's Time 3 L2 English CVC productions, tested on the CVC English Model with vowel duration included as a variable. 250

Table A8.5. Naturalistic group's Time 4 L2 English CVC productions, tested on the CVC English Model with vowel duration included as a variable. 250

Table A8.6. Naturalistic group's Time 5 L2 English CVC productions, tested on the CVC English Model with vowel duration included as a variable 251

Table A8.7. Naturalistic group's Time 6 L2 English CVC productions, tested on the CVC English Model with vowel duration included as a variable 251

Table A8.8. Training group's Time 1 L2 English CV productions in response to Voice 1 /b, pV/ stimuli, tested on the CV English Model with vowel duration included as a variable 252

Table A8.9. Training group's Time 2 L2 English CV productions in response to Voice 1 /b, pV/ stimuli, tested on the CV English Model with vowel duration included as a variable 252

Table A9.1. Recognition of English production tokens by vowel tested against the CVC English Model trained and tested on native speaker English productions with vowel duration excluded as a variable 253

Table A9.2. ANOVA and Multivariate test results for naturalistic L2 English vowel production data excluding vowel duration as a variable. 253

Table A9.3. Mean percent correctly recognized vowel productions over time for the naturalistic group's L2 productions (top panel), in contrast to the trained group's L2 productions described in Chapters 4 and 5 (bottom panel). Vowel duration was excluded as a variable in both the CVC English and CV English Models. 254

Table A9.4. Summary of mean percent correctly recognized L2 English productions for each English vowel, contrasting naturalistic vowel learning study (from Times 1-6) with trained vowel learning results (from Times 1-2). Vowel duration excluded as a variable. 255

Table A9.5. Naturalistic group's Time 1 L2 English CVC productions, tested on the CVC English Model with vowel duration excluded as a variable....................256

Table A9.6. Naturalistic group's Time 2 L2 English CVC productions, tested on the CVC English Model with vowel duration excluded as a variable......................256

Table A9.7. Naturalistic group's Time 3 L2 English CVC productions, tested on the CVC English Model with vowel duration excluded as a variable......................257

Table A9.8. Naturalistic group's Time 4 L2 English CVC productions, tested on the CVC English Model with vowel duration excluded as a variable......................257

Table A9.9. Naturalistic group's Time 5 L2 English CVC productions, tested on the CVC English Model with vowel duration excluded as a variable......................258

Table A9.10. Naturalistic group's Time 6 L2 English CVC productions, tested on the CVC English Model with vowel duration excluded as a variable......................258

Table A9.11. Training group's Time 1 L2 English CV productions in response to Voice 1 /b, pV/ stimuli, tested on the CV English Model with vowel duration excluded as a variable......................259

Table A9.12. Training group's Time 2 L2 English CV productions in response to Voice 1 /b, pV/ stimuli, tested on the CV English Model with vowel duration excluded as a variable......................259

List of Figures

Figure 1.1. Hypothetical distributions of competing L1 and L2 categories. 13

Figure 2.1. Crosslinguistic similarity continuum illustrating 'Same', 'Similar' and 'New' distinctions.. 30

Figure 2.2. Hypothetical L1/L2 category interactions illustrating the overlap between categories that are the 'Same', 'Similar' and 'New'. 31

Figure 2.3. A female native English speaker's [ɪ] production. 35

Figure 2.4. F1 and F2 formant values from the beginnings and endings of Canadian English vowels as well as their trajectory indicated by the arrows 36

Figure 3.1. Spectrogram and waveform of English [i]... 57

Figure 3.2. Eight alternative results of LPC automatic formant tracking based on frequency cut-offs between 3000 Hz and 4500 Hz (Successful first-pass). 58

Figure 3.3. Eight alternative results of LPC automatic formant tracking based on frequency cut-offs between 3000 Hz and 4500 Hz (Unsuccessful first-pass). 59

Figure 3.4. Manually adjusted formant tracks from among alternatives......................... 60

Figure 4.1. Screenshot of training program 85

Figure 4.2. Pooled training groups' mean correct vowel identification scores on Voice 1 /b, pV/ stimuli, at Time 1 and 2.. 91

Figure 4.3. Pooled training groups' mean correct vowel identification scores on Voice 1 /g, kV/ stimuli, at Time 1 and 2.. 91

Figure 4.4. Pooled training groups' mean correct vowel identification scores on Voice 1 /z, sV/ stimuli, at Time 1 and 2 92

Figure 4.5. Pooled training groups' mean correct vowel identification scores on Voice 2 /b, pV/ stimuli, at Time 1 and 2.. 92

Figure 4.6. Pooled training groups' mean correct vowel identification scores on /b, pV/ stimuli by Voice 1 and 2 at Time 1... 94

Figure 4.7. Pooled training groups' mean correct vowel identification scores on /b, pV/ stimuli by Voice 1 and 2 at Time 2... 95

Figure 4.8. LVT group's mean correct vowel identification scores on lengthened-vowel training stimuli at Time 1 and Time 2 97

Figure 4.9. SVT group's mean correct vowel identification scores on natural vowel training stimuli at Time 1 and Time 2 97
Figure 4.10. DVT group's mean correct vowel identification scores on natural vowel training stimuli at Time 1 and Time 2 98
Figure 4.11. Pooled training groups' mean correct identification scores on natural versus lengthened vowel stimuli at Time 2........ 100
Figure 4.12. Pooled training groups' mean correct vowel identification scores on Voice 1 /b, pV/ stimuli, at Time 2 and delayed post-test........ 104
Figure 4.13. Pooled training groups' mean correct vowel identification scores on Voice 1 /g, kV/ stimuli, at Time 2 and delayed post-test........ 104
Figure 4.14. Pooled training groups' mean correct vowel identification scores on Voice 1 /z, sV/ stimuli, at Time 2 and delayed post-test 105
Figure 4.15. Pooled training groups' mean correct vowel identification scores on Voice 2 /b, pV/ stimuli, at Time 2 and delayed post-test........ 106
Figure 4.16. Pooled training groups' mean correct vowel identification scores on /b, pV/ stimuli by Voice 1 and 2 at Time 2........ 107
Figure 4.17. Pooled training groups' mean correct vowel identification scores on /b, pV/ stimuli by Voice 1 and 2 at delayed post-test........ 108
Figure 4.18. Pooled groups' mean correct vowel identification scores on natural vowel training stimuli at Time 2 and delayed post-test........ 109
Figure 4.19. Pooled groups' mean percent correct production scores over time in response to /b, pV/ stimuli produced by Voice 1 113
Figure 4.20. Pooled groups' mean percent correct production scores over time in response to /g, kV/ stimuli produced by Voice 1........ 113
Figure 4.21. Pooled groups' mean percent correct production scores over time in response to /z, sV/ stimuli produced by Voice 1 114
Figure 4.22. Pooled groups' mean percent correct production scores over time in response to /b, pV/ stimuli produced by Voice 2........ 115
Figure 4.23. Pooled groups' mean percent correct production scores at Time 1 in response to /b, pV/ stimuli produced by Voice 1 and Voice 2........ 116

Figure 4.24. Pooled groups' mean percent correct production scores at Time 2 in response to /b, pV/ stimuli produced by Voice 1 and Voice 2 117
Figure 4.25. Comparison of average vowel perceptual identification and production recognition scores in response to Voice 1 /b, pV/ stimuli at Time 1 118
Figure 4.26. Comparison of average vowel perceptual identification and production recognition scores in response to Voice 1 /b, pV/ stimuli at Time 2 119
Figure 4.27. Comparison of average vowel perceptual identification and production recognition scores in response to Voice 2 /b, pV/ stimuli at Time 1 119
Figure 4.28. Comparison of average vowel perceptual identification and production recognition scores in response to Voice 2 /b, pV/ stimuli at Time 2 120
Figure 4.29. F0 normalized production measures for responses to Voice 1 /u/ and /ʊ/ stimuli at Time 1 122
Figure 4.30. F0 normalized production measures for responses to Voice 1 /u/ and /ʊ/ stimuli at Time 2 122
Figure 4.31. F0 normalized production measures for responses to Voice 2 /u/ and /ʊ/ stimuli at Time 1 123
Figure 4.32. F0 normalized production measures for responses to Voice 2 /u/ and /ʊ/ stimuli at Time 2 123
Figure 4.33. F0 normalized production measures for responses to Voice 1 /ɒ/ and /ʌ/ stimuli at Time 1 126
Figure 4.34. F0 normalized production measures for responses to Voice 1 /ɒ/ and /ʌ/ stimuli at Time 2 126
Figure 4.35. F0 normalized production measures for responses to Voice 2 /ɒ/ and /ʌ/ stimuli at Time 1 127
Figure 4.36. F0 normalized production measures for responses to Voice 2 /ɒ/ and /ʌ/ stimuli at Time 2 127
Figure 6.1. Mean correct scores of untrained speakers' L2 English vowel productions. 168
Figure A5.1. Mean identification scores pooled across vowels for each Training Day by each Training Group 239

Chapter 1. Second language speech learning

Although nearly everyone acquires a first language (L1) with relative ease, success in learning a second language (L2) varies dramatically across individuals, with some ultimately reaching a higher level of proficiency than others. However, even among those exceptional adult L2 learners who have seemingly mastered most aspects of the target language (e.g., syntax, morphology, vocabulary, etc.) it is very rare that their L2 phonological system is convincingly nativelike. It is well established that L2 accent in adult learners is affected by age of learning, the interaction between L1and L2 phonological categories and the degree of experience the learner has with the L2. While research in the area of L2 speech perception and production has been steadily increasing, many questions remain only partially answered. For example, what makes L2 phonological acquisition in adulthood so difficult compared to the apparent ease with which children learn their L1 phonological system? How might different approaches to training adult L2 learners result in improved perception and production of L2 sound contrasts? Are phonemic contrasts learned in the context of one syllable or word, transferable to new contexts? Do improvements in perception translate into improvements in production? What role does the learner's L1 play? How best can we measure differences between L1 and L2 phonological categories in terms of their phonetic similarity?

In this dissertation I seek to expand our current understanding of L2 speech perception and production. In particular, I review a number of claims concerning the nature of L2 speech perception and propose a model for more precisely measuring crosslinguistic vowel similarity. Next, I employ this model to assess the effect of training Mandarin L1 speakers to better perceive English vowel contrasts and to measure the effect of perceptual training on production; I also compare the model's predictions with the resulting L2 data set. Finally, I briefly contrast my findings with data from a longitudinal study of naturalistic English vowel learning. By addressing both conceptual issues as well as training in this dissertation, I attempt to build a bridge between theory and practice – something too often overlooked.

1.1. Fundamental differences in L1 and L2 speech learning

1.1.1. Critical Period Hypothesis

The inability of most adult L2 learners to develop an L2 phonological system that ultimately mirrors that of native speakers continues to demand greater explanation. Early on, much of the debate revolved around the Critical Period Hypothesis (CPH), originally proposed by Lenneberg (1967), in terms of first language acquisition. Lenneberg theorized that if a child did not learn his/her L1 by a certain age (around the onset of puberty), biological factors would prevent later acquisition. For many, the CPH offered an appealing explanation for the clear differences in success rates in second language acquisition (SLA) when contrasting younger and older groups of learners. Nowhere are these age-related differences more evident than in the area of L2 phonology. Indeed, the clear inability of most late learners to rid themselves of a foreign accent is the most-often cited evidence of a critical period for SLA (Patkowski, 1990; Scovel, 1969, 1988). Scovel (1969) maintains that because of biological constraints, particularly in relation to neuromuscular coordination, most post-puberty L2 learners retain a detectable foreign accent. For advocates of a critical period, those few learners who are successful in acquiring L2 phonology are simply exceptional cases for whom these constraints have somehow failed to apply. According to Gass (1984), adult L2 learners who appear to partially acquire L2 phonology do not do so through speech-specific mechanisms available to infants, which she maintains are no longer available, but instead, through reliance on general auditory mechanisms.

In contrast, the inability to explain exceptional learners' apparent immunity to supposedly normal biological processes has made many wary of applying an extreme view of the CPH to SLA. More moderate explanations have thus been postulated. Birdsong (1992), for example, argues that a cognitive critical period, rather than a biological one, offers a more appealing explanation for successful learners. For him, successful learners have maintained a degree of general cognitive flexibility unavailable to most learners – an ability he argues is evident in cognitive domains other than language.

Most arguments against a critical period are supported by research investigating just such successful L2 learners (e.g., Birdsong, 2007; Bongaerts, van Summeren, Planken, and Schils, 1997). For these opponents of a biological critical period account, the existence of even one exceptional learner is evidence enough to falsify the CPH. Rather than seeing successful learners as random linguistic deviants (in the same class as math geniuses and musical savants, etc.), they rely heavily on such cases to support their opposition to the CPH. This is not to say that opponents of a critical period deny that most adult L2 learners are unlikely to acquire nativelike L2 phonology. They simply deny that these challenges are biologically or cognitively insurmountable. Assuming a primarily perceptual root for accent, Bongaerts et al. (1997) conclude that there are only two reasonable possibilities when it comes to L2 speech learning by adults. Either original perceptual abilities that are available to children are no longer available, or, they are available, but accessing them is difficult. They conclude, on the basis of research showing that some adult L2 learners can develop accent-free speech, that perceptual abilities do indeed remain intact. To account for clear age-related effects for L2 accent, they argue that the difficulty many adult learners face stems from adult learners' general tendency to over-rely on L1 categorical perceptual strategies, rather than the continuous mode of perception used by children. This over-reliance on categorical perception is the result of adults ceasing to require the continuous mode of perception in their L1, once L1 phonological categories have been established. If adults are able to revert to the continuous mode of perception, they can then begin to establish new categorical boundaries for the L2. The ability to revert to the continuous mode of perception under ideal learning conditions has been demonstrated for the perception of slight differences in Voice Onset Time (VOT) by Kewley-Port, Watson and Foyle (1988). Since VOT is generally understood to be one of the most categorically perceived phonetic distinctions, the ability to revert to a continuous mode of perception in this context provides particularly strong evidence that this mode of perception is still available to many adults.

Studies of less successful learners also provide evidence against a biological critical period for the acquisition of L2 phonology. To date, Flege, Munro and MacKay's (1995) research provides some of the strongest evidence against the CPH. In a large study involving over 200 Italian immigrants to Canada, they found no evidence of a sudden loss or decrease in ability to produce unaccented L2 speech at a critical age boundary. Instead, they discovered a linear decline that started very early (about 4 years of age) and continued into adulthood. A similar linear relationship between what Munro and Mann (2005) term 'age of immersion' in the L2 and perceived accent has also been found to exist for Mandarin learners of English.

Some critics of both biologically and cognitively motivated critical periods for SLA argue in favor of social and psychological explanations. Schumann (1975, 1979) hypothesized that apparent age-related effects were confounded with the learner's social and psychological distance from the target speech community. Applying his Acculturation Model, one could easily argue that age of arrival in the speech community affects the type of exposure and interaction available to the learner. Other factors such as differences in motivation have also been shown to lead to differences in performance across learners (Cenoz & Garcia Lecumberri, 1999). While individual differences of the sort just outlined may help account for many learner differences in ultimate attainment of L2 syntax, morphology, and vocabulary, there is less evidence that ideal learning conditions will result in nativelike pronunciation. Munro and Mann (2005) propose a possible explanation. They concluded from their study of Mandarin learners of English that while a variety of psychosocial and contextual factors contributed to accent, perceptual factors were also extremely influential. The nature of L2 phonology is clearly set apart from other domains of SLA. For more detailed overviews of how learner variables may affect degree of perceived accent see Scovel, (1988), Flege, Bohn and Jang (1997), Flege, Frieda and Nozawa (1997), Gottfried and Suiter (1997), and Piske, MacKay and Flege (2001).

In summary, it is increasingly clear that empirical evidence contradicts rather than supports the existence of a biological critical period for SLA. Nevertheless, there is an indisputable relationship between age and ultimate attainment in an L2. The rest of this chapter will explore this issue further, focusing on the obvious differences that exist between L1 and L2 acquisition and offering an alternative account of the relationship between age and degree of L2 accent.

1.1.2. L1 speech learning

During the first year of life, infants are able to discriminate all phonetic contrasts, from the many world languages so far investigated, regardless of their ambient language.

Only after 6 to 12 months do they begin to lose this ability, a period coinciding with the formation of phonological categories found in their L1 language environment (Gerken & Aslin, 2005; Jusczyk, 1997; Kuhl & Iverson, 1995; Polka and Bohn, 1996; Werker, 1995; Werker & Curtin, 2005). This loss of phonetic discrimination ability is found to apply to both consonants (Best, McRoberts, & Sithole, 1989; Werker & Tees, 1984) and vowels (Kuhl & Iverson, 1995; Polka and Werker, 1994; Werker, 1995). It should be noted, however, that this loss of ability does not appear to apply equally to all sound contrasts. Polka and Bohn (1996) argue that some non-native contrasts may be inherently easier to discriminate than others, giving them a longer period of immunity to the sort of categorical processing that causes other sound contrasts to become less perceptible. This may be particularly true of vowel contrasts, since vowels are perceived more continuously than are consonants (Polka & Bohn, 1996; Strange, 1995). In other words, phonetic differences within vowel categories are more readily discernable by adults than are differences within consonant categories. One example of adult learners' ability to discern new vowel contrasts comes from Polka (1995), who found that some German vowel contrasts were easily discriminated by adult L1 English speakers for whom these contrasts were new. However, while new vowel contrasts were discernable to adults in Polka's (1995) study, this fact does not imply that adults are able to perceive vowel differences to the extent infants can. In fact, Polka and Werker (1994) found some indication that infants' ability to perceive most vowels may actually decline earlier than infants' ability to perceive consonants. However, since adults appear to be generally better able to discern new vowel contrasts than consonant contrasts, this may suggest that residual perceptual abilities for vowels are maintained longer than for consonants. Despite some variation in the age at which different contrasts become imperceptible, by early childhood, it is clear that, in general terms, the ability to perceive most contrasts is severely diminished, and usually lost. For example, Werker and Tees (1984) found that children as young as four years old performed as poorly as adults when asked to discriminate between sound contrasts not found in their L1.

During their early months, one fundamental advantage infants have over older children and adults is less competition for cognitive resources; they are not distracted by other aspects of language such as semantic or syntactic processing. They have the luxury of perceiving sound on its own and experimenting with it through babbling. Apart from the cognitive differences found in infant versus adult phonetic learning, infants and young children also benefit from inherently superior input (Cross, 1977; Fernald & Morikawa, 1993; Liu, Kuhl, & Tsao, 2002; Murray, Johnson & Peters, 1990; Snow, 1977). First, caregiver speech typically consists of much shorter utterances than those normally directed to adult L2 learners, which may allow for a more detailed evaluation of input before attentional resources are diverted to successive components of the speech stream. Child-directed speech also tends to be vowel-rich, so that more of it may be phonetically salient than is likely the case in adult-directed speech. Mothers have been found to actively lengthen vowels by as much as 300% (Kuhl & Iverson, 1995). Kuhl and Iverson also found that adults perceive vowels extracted from child-directed speech as better instances of the vowel category than vowels found in adult-directed speech.

There is some debate about the universality of infant-directed speech. For example, Ochs and Schieffelin (1994) argue that the nature of adult-infant interactions is culturally constrained. They suggest that what is commonly reported in research related to English or similar ethnolinguistic contexts is not necessarily true for other cultures. Contrasting infants from English, Kaluli (a Papua New Guinean community) and Samoan cultures, Ochs and Schieffelin provide evidence that the extent of modifications made in infant-directed speech differs across these three groups. In the latter two cases, infant-directed speech is very limited relative to that found in English speaking environments. The authors argue that this difference is related to the way adults view infants in each culture. For example, in Kaluli culture, it is reported that since infants are incapable of understanding speech, adults see no reason to speak to them. Only after Kaluli children demonstrate linguistic ability (i.e., they start to produce the Kaluli words for ‘mother’ and ‘breast’) do adults begin speaking to them directly. In the Samoan culture, infants are viewed as possible interlocutors only after they begin to crawl, indicating a perceived relationship between mobility and a child’s transition from infant status to status as a more mature and communicatively capable interlocutor.

Ochs and Schiefflin (1994) indicate that although early infant-directed speech in the Kaluli and Samoan cultures may be limited, deliberate strategies are used in training young children to speak in both Kaluli and Samoan. Inaccuracies in child-productions are explicitly corrected in the case of Kaluli; in Samoan, appropriate utterances are modeled by adults to young children, who are expected to imitate these utterances in the context of social interaction. Evidence of these child-directed speech strategies for post-infancy children seems to weaken Ochs and Schiefflin's main thesis, that language modifications are unnecessary for successful language acquisition.

Finally, Ochs and Schiefflin's (1994) counter-evidence to the universality of infant-directed speech is primarily focused on the lack of modifications adults make to vocabulary and syntax. Little attention is given to possible modifications of phonetic input. Other research indicates that in at least some non-English speaking communities adults do employ phonetic modifications in infant-directed speech. Liu et al. (2002) found that Mandarin mothers tend to amplify important acoustic information by hyper-articulating vowels. In addition, the researchers found a significant correlation between a mother's infant-directed speech and her infant's resulting speech perception in other contexts. The babies who received more hyperarticulated vowel input were better able to perceive vowel contrasts in new contexts, providing evidence that phonetic modification does matter.

1.1.3. L2 speech learning

L2 speech learning is fundamentally different from L1 speech learning in many respects. Although L2 classrooms often attempt to recreate ideal learning conditions, it is clear that no SLA context can recreate one crucial aspect of L1 learning: a blank slate. L2 learners already have a linguistic system in place. It is perhaps this one difference that singularly precludes the possibility of adults, particularly older adults, having a realistic chance of developing a native speaker-like L2 system. Another previously noted difference between naturalistic L1 and L2 learning is the degree to which other cognitive demands compete with phonetic learning for limited resources. Although a blank slate cannot be simulated in SLA, minimizing interference from other competing cognitive demands is possible, though not to the same extent as is the case for infants. In naturalistic L2 learning contexts, other cognitive demands likely influence the extent of acquisition. Lee, Cadierno, Glass, & VanPatten (1997), Schmidt (2001) and Van Patten (1996) all conclude that in naturalistic settings, meaning is attended to first, before form.

While their research deals primarily with form at the level of syntax, the same case has been made for phonetic learning. Flege (1995) argues that adult L2 learners fail to notice distinguishing properties of L2 sounds during on-line processing, but not in some other more favorable contexts where attentional resources are in less demand. Evidence of this is provided by Borden, Gerber and Milsark (1983) who found that learner imitation of nonsense syllables resulted in better pronunciation than did a real-word imitation task. This suggests an effect of lexical activation during which pronunciation previously associated with particular lexical items is automatically retrieved and cannot be easily suppressed.

Since L1 appears to be the single largest contributor to L2 accent, in the rest of this section, I will expand on the role of learners' L1 in L2 speech learning and review literature pertaining to attempts to better orient learners' attention to important phonetic cues in the L2 input through controlled exposure.

1.2. The role of learners' L1 in L2 speech learning

The notion that each L2 learner's L1 has a significant influence on SLA is not new, although our understanding of it has clearly evolved. Early attempts at describing L1 effects focused on errors in the L2 that were obviously related to patterns found in the learner's L1. Following Lado's (1957) Contrastive Analysis Hypothesis (CAH), explicit predictions were made regarding error patterns for a given group of L2 learners by comparing their L1 and L2 phonological systems. Lado argued that the degree of similarity between L1 and L2 systems is directly correlated to the degree of success and failure learners experience in acquiring the L2. In brief, if a phonological category in the L2 corresponds to a phonological category in the L1, positive transfer from the L1 will occur; if an L2 category does not have a corresponding category in the L1, negative transfer from a similar but not identical category will result. It was soon discovered that the CAH was inadequate in its ability to account for the entire range of errors present in the interlanguage of L2 learners. Corder (1971) and Selinker (1972) both recognized that learner errors could be defined as stemming not only from the L1, but that many are also developmental in nature, in much the same way that children's L1 speech contains developmental errors. As with L1 learning, some new L2 categories are said to be universally easier to learn than others. Eckman's (1977) Markedness Differential Hypothesis builds upon this claim, arguing that the degree of difficulty experienced in learning an L2 category stems not only from first language transfer, but is affected by the new category's degree of markedness relative to L1 systems. More recently, Major (1996) hypothesized that categories which are least similar to L1 categories, are more quickly learned than those that are more similar, because dissimilar sounds are immediately more salient. Similar claims regarding the effect of crosslinguistic similarity are posited by Flege (1995).

For more than a decade Flege's Speech Learning Model (SLM) has been the most comprehensive and influential model of L2 speech perception and production. It has been extensively developed and supported through a research agenda based on a set of explicit hypotheses concerning the nature of the developing L2 system and its interaction with L1 categories. Further details of the SLM will be provided later in this chapter.

Flege's research program as a whole has provided us with a much greater understanding of processes involved in developing L2 speech perception and production. Having established that biological factors associated with age are not likely causes of accent, (Flege et al., 1995), Flege and colleagues' research indicates that L2 learners' relative degree of experience with the L1 versus the L2 may be the single largest contributor to degree of accentedness. Since the relative amount and type of experience L2 learners receive is strongly confounded with age (i.e., the older one is, the more experience one has had with one's L1), it is unsurprising to observe correlations between accent and age. The amount of experience learners have with the L2 may explain much of why, on average, learners who begin using their L2 earlier in life tend to achieve more nativelike pronunciation than those who begin later: less experience with L1 categories in younger learners means less reinforcement of those categories. Theoretically, this suggests that for younger L2 learners, less experience with the L2 is necessary to overcome negative transfer from the L1 to the L2 system. As such, this provides a possible explanation for why younger L2 learners' degree of accent tends to be weaker than older L2 learners' degree of accent.

The impact that the overall quantity of exposure to the L2 has on ultimate attainment has been well documented. Yamada (1995), for example, found that Japanese speakers who had lived in the US for one year or more were better able to discriminate between English /l/ and /r/ than those who had arrived more recently. Although the more experienced Japanese learners of English in their study were not nativelike in their discrimination ability, they had begun to use nativelike spectral cues, while the inexperienced learners had not. Further evidence is provided by Flege, Bohn, and Jang (1997), who found significant accent rating differences when comparing experienced (*M* = 7.28 years in the L2 environment) and inexperienced L2 learners (*M* = .68 years in L2 community) regardless of L1 background.

In a follow-up to the Flege et al. (1995) study, Flege, Frieda and Nozawa (1997) grouped Italian L2 English speakers by their age of arrival (AOA) in the L2 environment to control for age and then looked for within AOA differences. They found that even within AOA groups, variation in accent was highly correlated with the continued use of the L1 vis-à-vis use of the L2. Most striking was a finding that even among those Italian immigrants to Canada who had begun learning English prior to the age of six, their ratio of L1 to L2 use had a demonstrable effect on their ultimate degree of accent in adulthood. That is, those immigrants matched by AOA who continued to use their L1 the most were found to maintain a more detectable L2 accent than those who primarily relied on their English L2 in daily interactions. These findings led Flege et al. (1997) to conclude that bilinguals have a single phonological system, made up of both first and second language categories. Consequently, many categories are in conflict. Those that are used most are stronger, suggesting that the more an L1 falls into disuse, the more nativelike the L2 categories can become. This suggests that rather than being constrained by a biological critical period, individual differences in SLA can be explained largely on the basis of competition between L1 and L2 categories in a single phonological system, something that is naturally correlated with age.

While such research provides convincing evidence that the relative use of the L1 and L2 constrains ultimate attainment, it should also be recognized that the absolute degree of experience necessary for an adult learner to become less accented likely varies in relation to the learners' L1. That is, interactions between particular L1s and particular L2s determine the amount of experience necessary to move toward more nativelike representations. L2 learners from L1s that are globally more congruent with the L2 may require less exposure because more positive transfer may occur from their L1 than is the case for learners whose L1s are globally more distinct. When learners can transfer some features of their L1 to the L2 (e.g., lexis or grammar), more of their attention can be devoted to those L2 features that are distinct, such as phonetic differences.

Evidence that degree of similarity between a learner's L1 and L2 predicts ultimate attainment is provided in studies such as Bongaerts, van Summeren, Planken, and Schils (1997) and Flege, Bohn, and Jang (1997). Bongaerts et al. (1997) found that under ideal conditions, late Dutch learners of English were able to speak without a detectable accent. They noted that Dutch is typologically very similar to English, giving this group of learners an advantage in that they experienced less negative transfer from their L1 system. By typologically similar, the authors seem to mean that the phonemic inventory as a whole has a large degree of overlap, although it is uncertain if this was determined impressionistically or on the basis of a quantitative acoustic metric. In any case, Dutch learners of ESL are able to borrow many categories from Dutch for use in English. Because these categories are so similar to their English counterparts, they suffice as substitutes and are perceived by native English speakers to be relatively unaccented. Presumably, this is in contrast to L2 speakers whose L1s contain phonological categories that although similar, are less so and therefore, when substituted for L2 categories are perceived as being more accented. For example, while there may be no difference in the learning processes Dutch L1 and Mandarin L1 English learners employ, native English speakers' impressions of Dutch versus Mandarin accented English differ substantially.

Another example of this phenomenon is reported in Flege et al. (1997) who found that German, Spanish, Mandarin and Korean learners of English performed differentially in developing English phonological categories. Where similar phonemic contrasts existed in the learner's L1, these categories transferred to the learners' English L2; new contrasts were more difficult to acquire. McAllister (2001) found similar L1 transfer effects for English learners of Swedish L2 vowel contrasts. Moyer (1999) found that of the 24 learners of L2 German in her study, only one was rated as having a nativelike accent by native speaker raters. She attributed the majority's apparent lack of success to factors such as degree of motivation, and quantity and type of instruction. However, Moyer's participants came from a variety of L1 backgrounds, most of which were unrelated to

German. Given the overwhelming failure of Moyer's participants to attain a nativelike accent in German, while under similar conditions Bongaerts et al.'s (1997) Dutch L1 participants were largely successful in acquiring L2 English phonology, it seems that Moyer should have given greater consideration to L1/L2 interactions in her analysis.

Although it appears clear that L2 categories can improve with greater exposure, caution should be exercised so as not to oversimplify the effect of exposure by treating all exposure as equal. Simply quantifying exposure in terms of months or years lacks precision and may lead to incorrect conclusions. For example, McAllister (2001) concludes that the role of experience is exaggerated, pointing to a weak correlation between length of residence (LOR) and success in L2 phonological attainment in his study of English L1 learners of Swedish. A possible reason he found such a weak correlation is that LOR does not usually provide a meaningful measure of quantity or quality of exposure to the L2. All learners with similar LORs do not share identical experiences with the L2 speech community. McAllister himself acknowledges that accurate measurement of exposure to L2 input is difficult.

Even in cases where the amount of exposure a learner receives is more quantifiable, individual tokens within that input are not all of the same quality. Each instantiation of an L2 category may differ in terms of its usefulness for L2 category formation or strengthening. Despite variation in the quality of input, frequency-based or exemplar accounts of category formation (e.g., Bybee, 2001; Pierrehumbert, 2001) appear to treat all within-category input as equal. It is not at all certain, however, given the complex interactions between L2 categories and pre-existing L1 categories, that all exposure is equally accessible as input to the new system. While frequency of a particular category may provide a general indication of the amount of input available, it seems reasonable to assume that not all instances of an L2 category are equal in the degree to which they are affected by the competing L1 category. Some may be perceived as good members of the L1 category, while others may be perceived as poorer members of that category. Similar reasoning has been proposed by Best (1995) in her Perceptual Assimilation Model (PAM). Figure 1.1 provides an illustration of this issue. Imagining hypothetical distributional properties for an L1 and L2 such as those illustrated, we might assume that instances of the L2 /ɪ/ that fall outside of the distributional properties of L1

/i/, though perhaps still perceived as L1 /i/, have a better chance of being perceived as different from the L1 than those L2 /ɪ/ productions that fall within the distributional properties of L1 /i/. Applying this view to a frequency-based account of category learning, we might conclude that it is not the absolute number of L2 /ɪ/ tokens that matters, but the number that are actually perceived as being poor members of the L1 category to which they are most likely to assimilate. Following this assumption, we must then conclude that any instances of L2 /ɪ/ that assimilate to the L1 /i/ category, but are not perceived as being unusual members of it, at best do not contribute to the formation of a new /ɪ/ categories. At worst, such instances of L2 /ɪ/ may strengthen faulty phonological representations that are predisposed to treat L2 /i/ and /ɪ/ as equal.

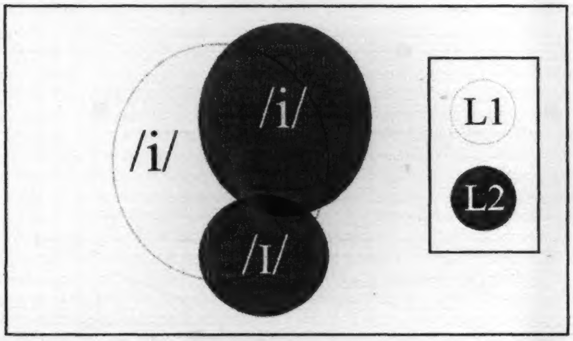


Figure 1.1. Hypothetical distributions of competing L1 and L2 categories.

This concept of what I will henceforth refer to as goodness-of-fit to the L1 category has been confirmed by Guion, Flege, Akahane-Yamada, and Pruitt (2000) and Aoyama, Flege, Guion, Akahane-Yamada, and Yamada (2004). In Guion et al., it was determined that the relative degree to which seven English and five Japanese consonants fit the opposing language categories differed. These differences were able to largely predict L2 confusion patterns, though not perfectly. Goodness-of-fit to the opposing language category was determined by asking native speakers of Japanese and English to categorize productions of opposing language consonants in /Ca/ frames in terms of their own language and then to provide a goodness rating. No crosslinguistic distinction was made between the vowel in these syllables in each language; /a/ was used to transcribe both English and Japanese. In the second study Aoyama et al. (2004) found that Japanese children living in the United States had less success acquiring English /l/ than /ɹ/, two categories that assimilate to a single Japanese category /ɾ/. They attributed this difference to the fact that these two English categories do not equally assimilate to the Japanese /ɾ/ category. English /ɹ/ is generally identified by Japanese listeners as less similar to Japanese /ɾ/ than is English /l/. Consequently, Japanese learners of English are better able to perceive acoustic information necessary for developing English /ɹ/. In contrast, because they perceive English /l/ to be virtually identical to Japanese /ɾ/, they have little motivation for developing the new English /l/ category and instead substitute Japanese /ɾ/ wherever English /l/ is required. Unfortunately, for English listeners, this substitution is perceived as accented. One limitation of this approach to measuring the effect of crosslinguistic similarity on L2 speech learning is that by averaging similarity scores across tokens, phonemic categories are treated as monolithic wholes, rather than as a set of independent instantiations, each of which is more difficult or less difficult to discriminate from L2 categories. Treating L2 tokens individually may provide more robust results. For example, it may be the case that one token of English /ɹ/ is more similar to Japanese /ɾ/ than is another token of English /ɹ/. The one that is more similar may be less discernable than the one that is less similar – something average similarity scores do not account for.

One explanation of why some tokens are more readily assimilable than others is elucidated in research by Kuhl and colleagues (e.g., Kuhl & Iverson, 1995). They have argued that as L1 categories become established in early childhood, they have a *perceptual magnet effect* on successive instances of the L1 category. In their account, the perceptual space closest to the categorical centre tends to warp future perception such that measurable acoustic differences between two stimuli are imperceptible if they fall within the space most influenced by what Kuhl and Iverson call the native language magnet. Yet, further from the categorical centre, similar acoustic differences between pairs of stimuli are perceptible, although both members of the pair are correctly recognized as members of the same category. While Kuhl and colleagues argue that a perceptual magnet attracts productions near category centres, one could as easily conclude that it is not categorical centres that warp perception, but rather that discrimination is enhanced at category boundaries because tokens in such locations are non-prototypical (e.g., Guenther, 2000). Whatever the case, it is clear from Kuhl and colleagues' research that near category centres, perception tends to be categorical, while near category boundaries perception is more continuous. In this dissertation I will continue to refer to this phenomena as a perceptual magnet effect, understanding as I have just indicated that the precise nature of the phenomena is uncertain. This warping of L1 perceptual space has serious implications for the processing of L2 sounds, particularly when a boundary between contrasting L2 categories falls within reach of an L1 perceptual magnet. The development of perceptual magnets for L1 categories through repeated experience may ultimately explain the perceptual difficulties adult L2 learners face. Older L2 learners may have stronger perceptual magnets than do younger learners.

1.3. Current models of L2 speech perception

Insights gleaned from the studies such as those discussed in the previous section have led to an increasing acceptance that limitations on ultimate attainment in L2 phonology are largely perceptual in nature, and directly affected by L1 and experience. Perceptual categories in the L1 have been divided up and reinforced in such a way that establishing new L2 categories at later ages is unavoidably problematic. The two most frequently cited models of L2 speech perception are Flege's (1995) Speech Learning Model (SLM) and Best's (1995) Perceptual Assimilation Model (PAM). In both models it is claimed that during the initial stages of L2 learning, most, though not all sounds in an L2 will be identified as an L1 category. The SLM and PAM both also make explicit claims predicting that success in acquiring L2 categories will vary in relation to each L2 category's interaction with the learner's L1 phonological system. The major difference between these two models is in focus. PAM is a static model of crosslinguistic speech perception, primarily addressing how categories in one language are perceived by monolingual speakers of a different language. Thus, this model attempts to identify crosslinguistic speech perception patterns for naïve listeners of the L2 (Best & Tyler, 2007). PAM is limited, then, to claims that can only apply to the initial state of L2 learners, before they have begun to develop L2 categories. Flege's SLM, on the other hand, describes dynamic patterns of L2 phonological learnability. SLM makes predictions in terms of which L2 categories will be easier to learn and which will be more difficult in relation to their interaction with L1 categories. As such, its focus is on ultimate attainability of L2 categories by advanced L2 learners. Both models are of importance to this study and will be briefly summarized in turn.

Perceptual Assimilation Model

Best's PAM (1995) posits three ways L2 categories can be perceived by the L1 phonological system. The first is through direct category assimilation, whereby an L2 category is heard as a member of a specific L1 category. In some cases two L2 categories are perceived as belonging to a single L1 category, characterized in PAM as "single-category assimilation." The second way an L2 category can be perceived is through assimilation to the L1 system as an uncategorizable speech sound. That is, it is heard as speech, but is not recognized as an obvious member of any L1 category. The third way an L2 category can be perceived is as a non-speech sound. In this presumably rare case, it is not assimilated to the phonological system at all.

Speech Learning Model

The SLM (Flege, 1995) makes predictions concerning the learnability of L2 categories in relation to their degree of similarity to existing L1 categories. A specific distinction is made between those L2 categories that are perceptually 'similar' to categories in the L2 and those that are dissimilar, or 'new' (Bohn & Flege, 1992; Flege, 1995). For late adult learners, 'new' L2 categories will be easier to learn and ultimately be more nativelike than 'similar' categories. In contrast, poor performance on 'similar' categories results from assimilation of L2 categories to 'similar' L1 categories (Flege, 1987, 1995). In general terms, the SLM notions of 'similar' and 'new' in terms of learnabilty correspond to PAM's static distinctions between those sounds that undergo direct assimilation and those that are uncategorizable speech.

Examples of PAM and SLM applied to speech learning processes

To illustrate how the predictions of PAM and SLM pertain to the issue of L2 speech perception and production, consider an example from L1 Spanish speakers learning English. In their L2 English production, English /i/ and /ɪ/ are often conflated into a single /i/ category. According to PAM, such an error stems from English /ɪ/ and English /i/ both undergoing direct perceptual assimilation to a single Spanish /i/ category at the beginning of the English learning process. However, PAM would likely indicate that both are not equally good members of the L1 Spanish category, something it terms ‘category goodness’. In fact, it has been quantitatively demonstrated that Western Canadian English /i/ and /ɪ/ are not equally good members of Spanish /i/. Morrison (2006) found that in terms of spectral properties, although Spanish /i/ is similar to both English /i/, and /ɪ/, Spanish /i/ is far closer to English /i/ than it is to English /ɪ/. However, the duration of Spanish /i/ is similar to that of English /ɪ/. According to SLM, English /i/ and /ɪ/ would both be deemed ‘similar’ to Spanish /i/ although as with PAM, not necessarily to the same extent. Given these assimilation patterns, PAM predicts initial difficulty in discriminating between English /i/-/ɪ/. From this point, PAM no longer makes any claims. SLM, however, predicts that differences in degree of similarity between English /ɪ/ and English /i/ might ultimately result in differences in performance at later stages of acquisition. If English /ɪ/ is less ‘similar’ to Spanish /i/, while English /i/ is more ‘similar’, SLM predicts a greater likelihood that Spanish L1 English learners will eventually perceive and produce English /ɪ/ as an independent category, while continuing to perceive and produce English /i/ as Spanish /i/. Although the overall likelihood of ultimate success may be greater for English /ɪ/ than English /i/, it is still possible that both members of the English /ɪ/-/i/ contrast are so similar to the L1 category that neither will be successfully learned and the learner will go on perceiving and producing both English /i/ and /ɪ/ as Spanish /i/. In this case, when English /ɪ/ is required, the intelligibility of Spanish accented English will be compromised, because Spanish /i/ is almost always perceived as English /i/ by NS English listeners.

One interesting example of ‘new’ categories reportedly being easier to learn comes from a study by Polka (1995) who found that English L1 adults had no difficulty discriminating between German vowel contrasts /u/ vs /y/ or /ʊ/ vs /ʏ/ without training. However, there is other evidence that discrimination of ‘new’ categories may not necessarily be easier at the beginning of learning or even at later stages. Guion et al. (2000) provide such conflicting evidence. In their study of Japanese learners of English, they examined the learners' ability to discriminate English-Japanese contrasts that comprised English sounds that were more or less similar to Japanese sounds. They found that although English /θ/ is relatively dissimilar from Japanese /s/, advanced proficiency adult learners from Japan demonstrated the same difficulty discriminating these crosslinguistic contrasts as did beginning proficiency adult learners. The same finding was evident for the English /l/-Japanese /ɾ/ contrast. The researchers did find a learning effect for the dissimilar English /ɹ/-/ɾ/ Japanese pair, however. Although Guion et al. conclude that these results only partially support SLM predictions, it is possible that the crosslinguistic contrasts examined, although dissimilar, were still similar enough that learning could not occur, or would require much greater exposure.

There is also evidence from L2 error patterns that all 'new' categories are not learned at the same rate or to the same degree. Munro and Derwing (2007) found that for both Mandarin and Slavic learners of English contrasts, some English vowel contrasts that the SLM would likely define as 'new' (e.g., /ɪ/ and /ɛ/), continue to be problematic after a year of L2 learning. One explanation may be that some contrasts are simply more difficult than others. Whatever the cause, SLM does not make predictions regarding the rate of acquisition of 'new' categories, but rather, only suggests that ultimately they will be more nativelike than 'similar' categories. Therefore, findings that some advanced learners have not yet learned 'new' categories do not necessarily falsify this prediction. It may be that more exposure is still needed, depending on both the nature of the L1/L2 interactions and L2 developmental patterns. Some L2 categories may simply be more difficult to learn than others and require greater exposure than other categories. In contrast, a finding that a 'similar' category was learned in a nativelike fashion, while a 'new' one was not, would create problems for the SLM. Another possibility is that the operationalization of 'new' versus 'similar' is not yet adequately defined, leading to faulty conclusions, something I will discuss in great detail in Chapter 2.

1.4. The relationship between perception and production in L2 speech learning

Before moving on to describe the effect of instruction on L2 speech perception, some mention should be made of the connection between L2 perception and L2 production. Since L2 learners' accent is defined in terms of their production capability, the link between the two is of utmost importance. Unfortunately, there is not a clear understanding as to how improvements in perception lead to improvements in production or vice versa. Certainly, one might expect that after learning to perceive L2 sound contrasts, a delay might be experienced before neurophysiological mechanisms can begin to implement L2 contrasts in production. Conversely, it might be possible to train learners on the basis of sound descriptions to produce sounds that they cannot yet perceive.

Gesturalist views such as Liberman and Mattingly's (1985) Motor Theory, and Fowler (1986) and Best's (1995) Direct Realist view posit a direct connection between perception and production. Although formulated somewhat differently, these views both hold that speech perception is achieved through reference to articulatory gestures, either directly or indirectly. For Best (1995), perception of sound categories is accomplished by referring to invariant properties extracted directly from distal articulatory gestures. For Liberman and Mattingly (1985), perception of categories from articulatory gestures is mediated through abstract phonological representations. In contrast, auditorist views such as those of Kingston and Diehl (1994) posit a primarily perceptual root for accent, with perceptual motivations underlying which gestures will be chosen for production. A more moderate middle ground is proposed by Nearey (1997), who argues for what he terms a 'double-weak' theory of speech perception. In brief, he believes that imperfect, weak connections exist between somewhat autonomous perceptual and gestural systems. Perception is informed by knowledge of the effect of gestural properties and selection of gestures is influenced by the perceptual tractability of the resulting sound. This is the approach most consistent with the SLM, given its claim that perceptual learning will eventually, if slowly, find its way into production, although mismatches may be present.

Unfortunately, very little research exists that explicitly demonstrates the nature of the L2 speech perception/production connection. Flege, Bohn and Jang (1997) do show that experience with the L2 has an effect on both perception and production, but not which comes first. However, it seems safe to assume, from studies of naturalistic L2 learning such as Flege et al. (1995) that learners must learn to perceive new sounds before they can begin to produce them. Otherwise, they would have no way of determining the extent to which their productions meet the L2 target. In the absence of explicit instruction, L2 speakers can only imitate the L2 in reference to their own perception. To account for Mandarin and Slavic learners developing English vowel pronunciation in Munro and Derwing (2007), the same explanation of perception before production is the most reasonable one. In the absence of significant explicit instruction in producing English vowels, these learners demonstrated improvement in pronunciation. Again, this can only be explained in reference to their own ability to perceive differences in the target language.

Following a gesturalist view, one could surmise that L2 perception and production must develop simultaneously. In fact, few studies indicated that perceptual training immediately transfers to production. Those studies that have addressed this issue provide contradictory evidence. For example Bradlow, Pisoni, Akahana-Yamada and Tohkura (1997) found that training Japanese L1 learners of English to perceive the /l/-/r/ contrast transferred to production. In contrast, Wang (2002) found that training Mandarin and Cantonese learners of English vowels to perceive target contrasts did not immediately transfer to production. One possible explanation for such contradictory results is that some contrasts may simply be more difficult to learn to produce than are others, increasing the lag time between perception and production. For example, the basic gestures underlying a contrast such as /l/-/r/ may have been understood by the Japanese learners in Bradlow et al.'s (1997) study because of previous English instruction prior to their perceptual training, or detected during it, aiding transfer between perception and production. In contrast, specifying vowel gestures and detecting them may be much more difficult, offering a possible explanation for why improvement in vowel production may not immediately follow perceptual training. Learners are unlikely to have previously learned about the shape of the vocal tract associated with particular vowels and detecting their shape in perception would thus be difficult without simultaneously practicing their production.

It should be noted that one study has also demonstrated that production can precede perception. Sheldon and Strange (1982) found that Japanese learners of English were able to learn to produce an English /l/-/ɹ/distinction without first being able to perceive this contrast. Given that the learners in this study were made aware of the gestural differences used to produce these sounds, it does not reflect the type of learning possible in most naturalistic contexts where explicit articulatory instruction is not often provided.

Whatever view one holds concerning the perception/production connection, there is a general consensus that it is the difficulty or inability to perceive phonetic contrasts that causes most phonetic inaccuracies in L2 production (Archibald, 1998; Flege, 1981, 1984, 1995; Rochet, 1995). Flege (1981) believes that with sufficient exposure improvement in production is achievable, pointing out that despite *increasing* ability in non-linguistic sensorimotor skills (e.g., the coordination necessary for sports), children start losing the ability to produce L2 vowels and consonants. He argues that this mismatch between worsening speech articulation ability and improving general sensorimotor ability makes a perceptual explanation for accent most appealing. He suggests, then, that it is not the learners' inability to control physiological speech mechanisms, but rather, their inability to perceive new categorical distinctions that causes L2 accent. If Flege is correct, learners should be able to train their physiological mechanisms to recreate those L2 sounds they accurately perceive.

1.5. Attention and instructional intervention in L2 speech learning

I will conclude this chapter with a discussion of the effects of instructional intervention. First, I will discuss why I believe instruction is important from a cognitive perspective. Then, I will review a number of studies that demonstrate that instructional intervention can improve on the results experienced by L2 learners in entirely naturalistic learning environments.

Given that degree of L2 experience has a demonstrable effect on ultimate attainment, it seems reasonable to conclude that the quality of that experience may also be important. We know from L1 research, such as that mentioned earlier, that modified speech by caregivers appears to facilitate L1 acquisition. Since the nature of L2 input in naturalistic settings is much different for adults than for children learning their L1 or even children learning an L2, we might expect this important difference could account for some of the difficulties adult learners of an L2 face. If the nature of the input available is a limiting factor, making L2 input more salient or noticeable, as it is for L1 learners, could have a facilitating effect on SLA. I noted earlier that Schmidt (2001) argues that in naturalistic L2 learning environments, adult learners usually attend to meaning to a greater extent than form. Similarly, I reported that Flege (1995) argues that adult L2 learners are unable to make use of important phonetic information in some learning contexts, but are able to when their attention is more directly focused on a phonemic discrimination task.

In earlier research (Thomson, 2003), I conclude that learning some new L2 phonetic contrasts may be especially difficult in naturalistic settings if only a small proportion of tokens of the target category are perceptually salient (i.e., most tokens are directly assimilated to an L1 category while only a few are noticed as being deviant members of all L1 categories). In such cases, the majority of available input for that category goes undetected. That is, during naturalistic processing of some L2 categories, many tokens or instances of a category are not salient enough to capture the learners' attention as being 'new'. Consequently, to develop and strengthen such categories, learners require greater input of that category as a whole than they do for categories where a larger proportion of tokens are clearly dissimilar from any L1 category. A major benefit of instruction, then, lies in its ability to orient learners' attention to important cues in the input that are not easily detectible in naturalistic learning environments. This has the effect of allowing them to make use of more of the input they receive.

The effect of attention in SLA domains other than phonology is well documented. Schmidt's (1990,1993, 2001) claim that 'noticing' or 'awareness' of L2 forms is a precursor to acquisition of those forms has been very influential in bringing the issue of attention to bear on our understanding of adult acquisition of L2 syntactic features. Others have reached similar conclusions. Tomlin and Villa (1994) argue for a multidimensional approach to attention, whereby three levels of the attentional process all have a potential impact on SLA. Tomlin and Villa's approach extends Posner & Peterson's (1990) description of the human attention system to second language acquisition. In this approach, attention is divided into three processes: alertness, orientation, and detection.

By alertness, Tomlin and Villa (1994) mean learners should be generally focused on the task at hand, not distracted by other cognitive demands. Orientation is the more specific directing of attention to the class of linguistic objects or stimuli of interest. Finally, detection is the process during which important differences in stimuli are actually detected as meaningful for successful communication. Unlike Schmidt's (1993), 'noticing' hypothesis, however, Tomlin and Villa (1994) maintain that learners need not be consciously aware of the detection process. Detection is the process most likely to result in real changes to the developing linguistic system, while alertness and orientation increase the probability of detection.

However, the positive effect of consciously orienting L2 learners' attention to critical information has been shown in classroom-based research examining different types of second language instruction (Doughty & Williams, 1998; Sharwood Smith, 1993; Van Patten 1996, 2002, White, 1998). Lyster and Ranta (1997) also found that certain types of corrective feedback during more naturalistic communicative tasks resulted in modification of the learners' developing system while others did not, suggesting that some forms of orienting attention are not as likely to result in detection as others. In general, when target forms are made more salient through the use of corrective feedback or other means, they are more likely to be detected and incorporated into the developing L2 grammar. Again, most research has been limited to domains other than phonology, and most often to the acquisition of L2 syntax and morphology. It should also be mentioned that research in other SLA domains has largely found that the rate of acquisition of certain morphological or syntactical features is often increased through instruction, though not the order in which features are acquired (Ellis, 1990; Long, 1983; Pica, 1985). This may have implications for phonetic learning as well, although I am unaware of any research concerning developmental sequences in L2 phonology.

The role of attention in L2 phonetic learning has been explicitly tested in research conducted by Guion and Pederson (2007a). In this study, two groups of English monolinguals were trained to perceive five Hindi stop consonant contrasts, three of which, /b/-/bʱ/, /t̪ʰ/-/ʈʰ/ and /k/-/g/[1], are particularly difficult for English speakers to discern. In training, all stop contrasts were presented in the context of a Hindi word learning task.

---

[1] In this contrast the /k/ is unaspirated.

One group of learners (sound-attending) was instructed to attend to the beginning (i.e., the first consonant) of each word and was explicitly alerted to the fact that some Hindi sounds that are lexically contrastive may often seem similar; learners in this group were asked to try to distinguish between such sounds as best as they could. A second group of learners (meaning-attending) was trained on the same Hindi word stimuli, but was only instructed to learn the meaning of each Hindi word; they were not alerted to the fact that in some word pairs with different meanings the initial sound may seem similar, nor were they instructed to pay particular attention to the beginning of the word. A comparison of the learners' pre and post-test scores on an ABX discrimination test indicated an effect of training condition. After training, the sound-attending group demonstrated improvement in their ability to discriminate the /t̪ʰ/-/ʈʰ/ contrast, while the meaning-attending group did not. Both groups, however, improved to the same extent in their ability to discriminate the /b/-/bʱ/ contrast. Finally, neither group improved in their ability to discern differences between /k/-/g/.

The fact that the sound-attending group was better able to learn the /t̪ʰ/-/ʈʰ/ contrast was interpreted by the researchers as evidence that orienting learners' attention to phonetic contrasts in a second language has a positive effect on the development of L2 speech perception. In contrast, although both training groups improved from pre-test to post-test in their ability to identify the meaning of Hindi words on a semantic test, the meaning-attending group improved to a larger extent than the sound-attending group. This difference was attributed to an effect of orienting the meaning-attending group's attention to the semantic aspect of the stimuli.

In another study (Guion and Pederson, 2007b), English speakers were trained on monosyllabic Hindi word contrasts varying by onset consonant and medial vowels. One training group was instructed to pay attention to the consonant in each syllable, while a second training group was instructed to pay attention to the vowel portion of each syllable. A comparison of pre and post-test results using a discrimination test indicated that the consonant-attending group improved in their ability to discriminate Hindi consonant contrasts, while the vowel-attending group did not. Neither group demonstrated improvement in their ability to identify Hindi vowels; however, this was predicted based on the fact that the learners' scores on the vowel discrimination pre-test were already near ceiling (96.8%). As with Guion and Pederson's (2007a) first study, the results of this study support claims that attentional orienting plays an important role in L2 phonetic learning.

Previous phonetic training experiments have also been shown to have positive effects on L2 phonological acquisition. While some studies demonstrating the effect of explicit pronunciation (perception and production) training have used natural training stimuli (e.g., Derwing, Munro & Wiebe, 1997, 1998), there is evidence that modified training stimuli may increase the degree of success learners experience. Jamieson and Morosan (1986) applied a perceptual fading technique in which exaggerated frication cues were used to help Francophones notice the English voiced/voiceless interdental fricative contrasts. They found that by initially enhancing the frication cues using synthetic training stimuli, the learners were better able to notice these L2 categories. After successful learning of these exaggerated L2 contrasts occurred, Jamieson and Morosan then reduced the degree of enhancement in equal steps until reaching more nativelike training sets. Learners in this study were generally able to maintain the contrast, though they made more errors in response to more natural stimuli than with the maximally enhanced stimuli. In a later study, Jamieson and Morosan (1989) found improvement in category perception also occurred when learners were trained on prototypical exemplars of the categories, rather than on the synthetically enhanced stimuli used in the Jamieson and Morosan (1986) perceptual fading experiment. Interestingly, significant improvement was detected after only 40 minutes of training, though all the participants had had extensive previous experience with English.

While prototypical exemplar training facilitated learning in the Jamieson and Moroson (1989) study, this type of training may not be as effective in all contexts. A crucial finding regarding instructed L2 speech learning using synthetic stimuli is that high variability training often results in better learning than training which only relies on best exemplars (Logan, Lively & Pisoni, 1991; Pisoni & Lively, 1995). In more recent research, Jongman and Wade (2007) provide evidence that for easier L2 distinctions, high variability training may be better, while for more ambiguous L2 contrasts, training on prototypes may be better. In the best-exemplar training paradigm, learners are trained on stimuli that are characterized as being the best examples or models of the target category. Usually, synthetic stimuli that possess the properties of average native speaker productions for that category are used. Stimuli are intended to be immediately recognizable by native speakers as instances of the target category. The problem for such a model is that it assumes that idealized native speaker norms will provide the best training for all non-native speakers. While such stimuli may provide a good model in terms of the intended target, it is not at all certain that important acoustic information in such stimuli is easily perceived by all learners from all L1s. Such a training paradigm does not take into account possible interactions with the learners' L1s or emerging L2 categories. A best-exemplar for a native speaker of an L2 may not be well perceived by L2 learners whose L1 perceptual magnets are near that exemplar. In contrast, for a learner from a different L1 background, the same best-exemplar may fall beyond the reach of any L1 categorical magnet affect and therefore less perceptual distortion is likely to occur. Following this reasoning, high variability training paradigms may work because they provide sufficient variation within a given category that presumably at least some tokens will be detected by the learner as members of a new category. Those production tokens that are noticed by a given learner are not necessarily prototypical instances of the category. Hence, it may not be variability itself that promotes learning, but the fact that the higher the variability, the better chance a sufficient number of tokens will be noticed by the learner. This also provides a possible explanation for why absolute frequency effects posited by Pierrehumbert (2001) for L1 phonological learning do not easily extend to L2 phonological learning. Despite similar amounts of exposure to L2 categories in absolute terms, the rate of acquisition of each L2 category varies. Pierrehumbert argues that an L1 phonological category reflects the sum all previously encountered tokens of that category. In other words, when a speech sound is heard, it is perceived as a specific phonological category through comparing its phonetic properties to previously experienced tokens of existing categories. Applying such an approach to the development of L2 phonological categories, we might expect that some tokens of a new category may be similar to tokens of an L1 category and therefore will be mislabeled as the L1 category. Other tokens may be different enough to justify labeling them as a new category against which future tokens can then be compared. Since at the beginning of L2 learning many

L2 tokens may be labeled as L1 categories, we cannot assume that absolute frequency matters, but rather only the frequency with which L2 productions are correctly identified as a new L2 category vis-à-vis the frequency with which they are incorrectly identified as a preexisting L1 category. Instructed learning has the potential of providing greater experience with correctly identifiable L2 tokens.

Another important variable in controlled phonetic learning contexts concerns the nature of feedback. McCandliss, Fiez, Protopas, Conway, & McClelland, (2002) and McClelland, Fiez, & McCandliss (2002) conducted experiments that applied adaptive phonetic training comparing two conditions: training with and without feedback on incorrect responses. Specialized phonetic training was provided that, like the Jamieson and Morosan (1986) study, involved artificially enhancing phonetic cues, this time in the context of Japanese learning to perceive English /l/ vs. /r/ categories. In McCandliss et al. (2002), the modified stimulus item was presented adaptively, meaning that when a stimulus was correctly identified by the learner, a less exaggerated version was presented next. If an incorrect response was made, a more exaggerated version followed. McCandliss et al. (2002) found that, given sufficient time, adaptive training had a positive effect on learning. However, when the training was coupled with feedback on correct versus incorrect responses, learning was more rapid. Most encouraging was the fact that this learning appeared to generalize to novel contexts, as was also accomplished in another study of Japanese learners of English /l/ and /r/ by Logan, Lively and Pisoni (1993).

Not all studies concerning the role of instruction have reached entirely positive conclusions. In one training study, Cenoz and Garcia Lecumberri (1999) provided their participants with identification training for English vowels. They found that while those who scored lowest on the identification pretest demonstrated a significant improvement in identification at the end of the training period, those who scored highest on the pretest did not improve. From this result, they speculate that those who performed best before training had reached a limit short of native-speaker ability. Such a conclusion does not bode well for those who believe that pedagogical intervention can help overcome the limitations imposed by other learner characteristics. However, Cenoz and Garcia Lecumberri (1999) do not describe the precise nature of their vowel training regimen.

They indicate that 14 hours of training were dedicated to improving English pronunciation, but this included training in consonants as well as prosodic features. The exact time allotted to training in English vowel identification is not specified. Furthermore, they state that they relied on commercially available English pronunciation material. Commercial material does not typically incorporate a large number of speakers, suggesting the training stimuli used in Cenoz and Garcia Lecumberri's study may have lacked sufficient variability. Given potential limitations in the training method the researchers used, their conclusion that some learners had reached a limit in their ability to identify English vowels remains open to debate. They do acknowledge that with longer training, or a different type of training, those whose learning curve appears to have slowed or plateaued may still be able to make additional progress.

1.6. Summary

In summary, difficulties learners face in the development of L2 speech perception is affected by degree of experience with the L2 as well as interactions between L1 and L2 speech categories. Creating an environment in which adult L2 learners' attention is more deliberately oriented to maximally salient input seems like a reasonable starting point for enhancing the effect of experience on developing L2 phonology. While limiting the effect of competition for cognitive resources is difficult, enhancing the input may be more readily achieved. Although adults learning a second language have strongly reinforced phonological categories that may have a perceptual magnet effect on new input, it has been demonstrated that sufficient experience can partially overcome this effect; new L2 categories do emerge over time. If a greater proportion of the input is made salient, and the learners' attention is drawn to important distinctions, the effect may be amplified. Essentially, pedagogical intervention affords the opportunity to provide a more effective L2 acquisition experience over a shorter period of time than is possible in naturalistic settings.

Chapter 2. Measuring crosslinguistic vowel similarity

Having established the important role that a learner's L1 plays in determining his or her ability to perceive L2 sound contrasts, in this chapter, I will discuss the issue of crosslinguistic similarity in greater detail, and propose a new statistical model for more effectively measuring it. Accurate measurement is important if the ultimate goal is to predict and understand interactions between L1 and L2 categories for specific L1s and specific L2s. A better understanding of these interactions may in turn allow us to more adequately determine what form(s) instructional intervention should take.

2.1. A crosslinguistic similarity continuum

As was mentioned in the previous chapter, Flege's (1995) SLM posits a binary distinction for comparing L2 sounds to the learner's L1 system. To briefly review, the SLM argues that 'similar' sound categories are more difficult to acquire in a nativelike fashion than are 'new' sound categories. In later versions, the SLM does recognize that there is some gradation in terms of how 'similar' L1 sound categories are to L2 sound categories (Flege, 2005). While many research findings support SLM's 'similar' versus 'new' distinction, the SLM fails to account for the sort of findings identified by Bongaerts et al. (1997), who claimed adult Dutch L1 learners of English were able to overcome accent in part because their L1 contains categories that are very close to English categories. Applying claims of the SLM, the similarity of Dutch to English should result in accented productions because learners are simply bootstrapping on their L1 categories.

For the purposes of measuring crosslinguistic similarity, I believe it is more accurate to conceive of L1/L2 contrasts along a crosslinguistic similarity continuum that includes a class of L1/L2 contrasts that I will refer to as the 'same', as illustrated in Figure 2.1. below. It is possible that some 'similar' L1 and L2 categories may be 'similar' to such an extreme that slight differences are undetectable to the naïve human listener. That is not to say that they are not quantitatively different, but they are unlikely to be reliably perceived as different to the average native speaker and are therefore readily substitutable for the L2 sound without affecting intelligibility or accentedness. This would give them a *de facto* 'same' status.

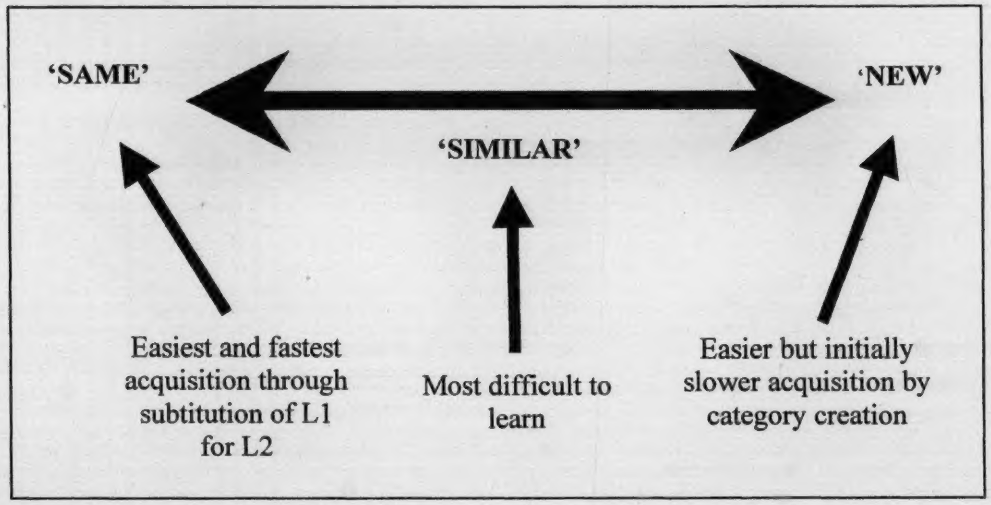
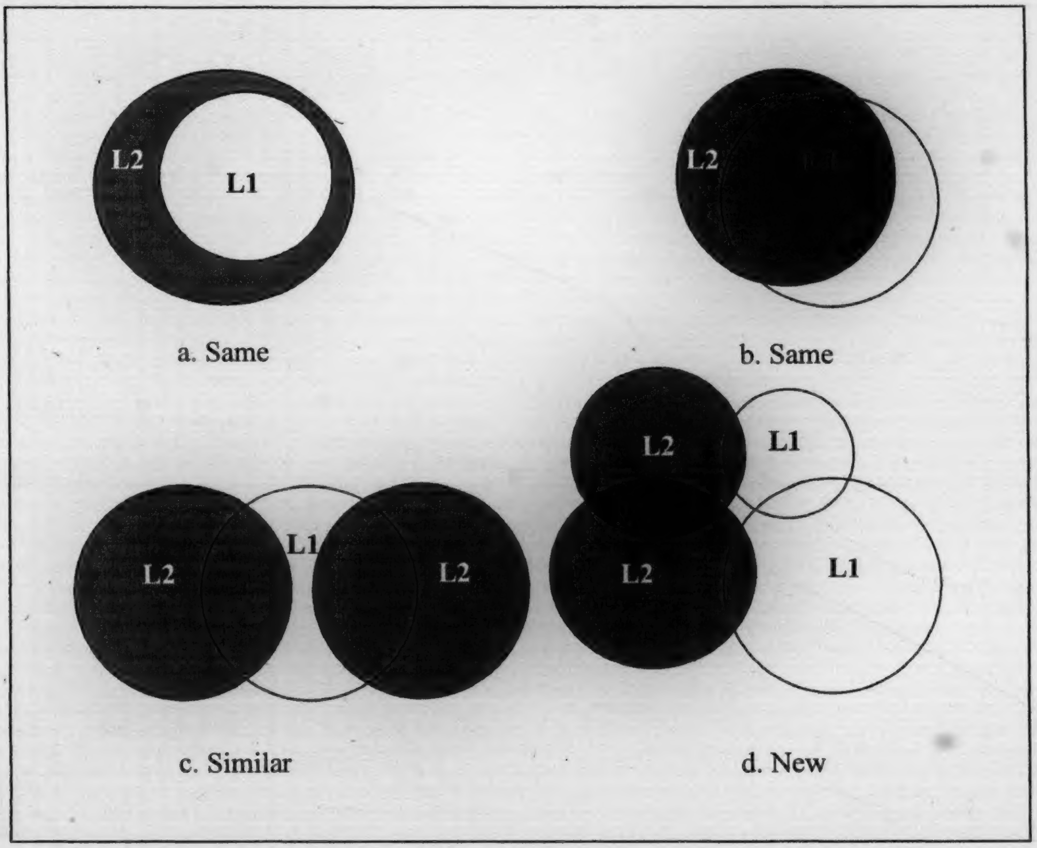

'SAME'
'NEW'
'SIMILAR'
Easiest and fastest acquisition through subtitution of L1 for L2
Most difficult to learn
Easier but initially slower acquisition by category creation

Figure 2.1. Crosslinguistic similarity continuum illustrating 'Same', 'Similar' and 'New' distinctions.

Several alternatives in Figure 2.2 expand on this three-way distinction in a simplified form; for the purpose of illustrating the basic concepts, other surrounding categories that may lead to confusions are temporarily ignored. If the distributions of competing L1 and L2 categories were found to be like those in Figure 2.2a, we might expect the L2 learner to have little difficulty perceiving the L2 category in contrast to other categories and find that the L2 speakers' productions of the L2 category are entirely unaccented, although the amount of variation in production would likely be smaller. If the distributions of L1 and L2 categories were closer to those represented in Figure 2.2b, we would expect there to be very little perceptible accent most of the time. On occasion, some substitutions of the L1 category for the similar L2 category may be from a part of the L1 category's distribution that falls outside of the typical L2 production range. In such cases, the L2 production may be perceived as accented by native speaker listeners, although they would not necessarily be unintelligible as the intended category.



a. Same

b. Same

c. Similar

d. New

Figure 2.2. Hypothetical L1/L2 category interactions illustrating the overlap between categories that are the 'Same', 'Similar' and 'New'.

The two types of 'same' categories illustrated in Figures 2.2a, and 2.2b would result in behavior that is quite different than we might expect if two L2 categories were 'similar' to a single L1 category and had distributional properties closer to those represented in Figure 2.2c. Substituting a member of the L1 category illustrated in Figure 2.2c in production of one of the two L2 categories illustrated would be more likely to result in the perception of accent. In cases when the speaker substitutes a production of the L1 category for one of the L2 categories, but from a portion of the L1 category's distribution that is closer to the unintended L2 category, the production may be unintelligible.

Finally, Figure 2.2d. illustrates a scenario where 'new' categories are quite distinct from any competing L1 category, although a few outlying members within the L1 and L2 distributions might be somewhat similar to the opposing language's category. If these L2 distributions are treated as 'new' categories, any difficulty in perception or production will most likely be the result of an interaction with other L2 categories in the developing system, not as a result of interactions with an L1 category. The hypothetical scenarios presented in Figure 2.2, demonstrate why it is more accurate to conceive of L1 and L2 categories as being more or less 'similar' with the qualification that when categories are nearly the same, they need not be learned as 'new' categories. Rather, there are contexts where the L1 category is so similar that it will immediately suffice in place of the L2 category.

A crucial distinction that I wish to emphasize at this point is that I conceive of phonological categories in terms of their distributional properties. More specifically, I explicitly recognize that L1 categories are made up of specific tokens, each with its own unique phonetic properties. For an L1 speaker, the phonetic variability that exists within a category is relatively unimportant as long as each token is identifiable as a member of the intended category, which is usually the case. However, when considering the acquisition of an L2 phonological category, phonetic variation in the input is of critical importance. That is, while a set of production tokens can be defined as a 'category' for L1 speakers, for L2 learners, especially beginners, each production token should be viewed as having a unique interaction with each learner's relatively stable L1 categories. Hence, when I refer to 'same', 'similar' and 'new', these terms should be taken to mean that most native speaker productions of the L2 category fall near 'same', 'similar' or 'new' points on the crosslinguistic similarity continuum, not that *all* tokens within each category are equally 'similar' or 'new'. The precise nature of each 'same', 'similar' and 'new' category is uncertain in the sort of crisp categorical terms that can be used to describe relatively stable L1 categories. For practical reasons, and as a means of containing discussion, I will continue to use the term 'category' in reference to both L1s and L2s. However, as will become evident throughout this dissertation, although clustering L2 productions into such classes is practically useful, I believe that a more precise way of envisioning crosslinguistic similarity is to view every production token in the L2 learner's input as being individually more or less similar to categories within the learner's L1 system.

From a theoretical perspective, 'same', 'similar', and 'new' distinctions seem intuitively appealing. Flege (1995) readily admits, however, that determining the degree of crosslinguistic similarity is inherently difficult since adequate operationalization of these terms continues to be elusive. After providing a brief overview of historical approaches to measuring crosslinguistic vowel similarity, I will propose a quantitative approach that offers a more precise way to operationalize and measure crosslinguistic similarity.

2.2. Past approaches to crosslinguistic similarity

Evidence for the general plausibility of a 'new' versus 'similar' distinction in L1/L2 speech categories has existed since the days of contrastive analysis (Lado, 1957), as discussed in the preceding chapter. To review briefly, in contrastive analysis, the determination of similarity starts with the assumption that any errors found in the L2 are the consequence of interference from the L1. This tradition continues to some extent to be applied in more recent discussions of crosslinguistic similarity. For example, finding that English nominal monophthongs are more difficult for Spanish and Basque speakers to acquire than are diphthongs, Cenoz and Garcia Lecumberri (1999) conclude that English diphthongs are more similar to competing L1 sounds than are monothongs, allowing positive transfer. No attention is given to the potential interaction with L2 developmental processes.

Comparing German, Spanish, Mandarin and Korean learners of English, Flege, Bohn and Jang (1997) found that Germans are quick to acquire the /i/-/ɪ/ contrast in English, while speakers of the other three languages are not. They attributed the German group's success to the fact that German has similar contrasts, while the other three languages do not. Similarly, Flege (1995) discusses an unpublished study of Korean L2 learner productions of English /i/ and /ɪ/ where he found that Korean-accented English /i/ productions were perceived by native English speakers as /ɪ/ 33% of the time, while the Korean-accented English /ɪ/ productions were perceived as /i/ 23% of the time by the same listeners. From such evidence, he argues that Koreans have a single /i/ category containing both English /i/ and /ɪ/.

While working backwards from L2 error patterns provides general support for the 'new' versus 'similar' distinction, it is limited to specific contrasts, and does not always explain the direction or magnitude of error patterns, neither does it clearly account for all errors. Claims concerning which L2 sounds are 'similar' and which are 'new' that rely on impressionistic category assignment also have limitations. For example, although a Mandarin /u/ may be heard by an English speaker as a member of the English /u/ category, it does not follow that the converse is also true. An English /u/ is not necessarily perceived as a Mandarin /u/ by Mandarin speakers. Hence, attempting to describe crosslinguistic phonetic similarity in terms of broad phonetic transcription of sounds may result in potentially inaccurate analyses.

To their credit, proponents of SLM recognize the need for a more precise means of defining 'similar' versus 'new' (Flege, 1995, 2005). Flege (1995) outlines a variety of metrics that have been used in the past. He is particularly critical of approaches such as those just described that rely on abstract phonological representations, where L2 sounds that are not found in L1 phonemic inventories are termed 'new' while L2 sounds with a counterpart in an L1 inventory are deemed 'similar'. He also discusses approaches that rely on general auditory properties associated with L1 and L2 sounds, using human listeners (as in Flege, Munro & Fox, 1994) and approaches based on putative articulatory gestures and vocal tract constrictions.

Within the SLM research paradigm, one of the most commonly used approaches for defining crosslinguistic similarity is to refer to shared spectral properties. For example, Flege (1995) reports that in one study he found that Spanish speakers' productions of English /i/ exhibited substantial spectral overlap, in an F1-F2 space, with native English /i/ and /ɪ/. This evidence was used to support claims that Spanish /i/ subsumes both English /i/ and /ɪ/. Likewise, Bohn and Flege (1992) compared English /i/, /ɪ/, and /ɛ/ with German /i/, /ɪ/, /ɛ/ and /ɛː/, having first predicted that these categories were similar on the basis of a review of descriptive studies. They then drew ellipses around distributions of these vowels produced by ten English and ten German native speakers, based on 95 % confidence levels for two principle components of variation, vowel height and frontness. On the basis of overlap between the resulting ellipses, Bohn and Flege conclude that the English vowel categories /i/ and /ɪ/, are 'similar' to German /i/ and /ɪ/ respectively, while English /ɛ/ is 'similar' to both German /ɛ/ and /ɛː/.

While much stronger than other methods in its objectivity, this method still lacks precision on a number of fronts. First, it is static, not taking into account spectral change that might occur within vowels. Assumptions made on the basis of so-called steady state portions of vowels are often inaccurate, as demonstrated by Nearey & Assmann (1986) who found that for English vowels, vowel inherent spectral change (VISC) is an important factor in accurate vowel identification. VISC refers to the gradual shift of formant frequencies over the length of the vowel, regardless of consonantal context. Hence, it seems invalid to conceive of English vowel perception purely in terms of so-called steady state portions of vowels. Since early research on VISC, similar patterns have been found for Michigan English vowels (Hillenbrand et al., 1995) and North Texan English vowels (Assmann & Katz, 2000).

VISC is illustrated in Figure 2.3. This spectrogram of a female English speaker's [ɪ], extracted from a [bɪ] syllable, shows substantial movement over time in the spectral dimension. In particular, notice that F1 is rising while F2 is falling, long after the transition from the preceding consonant into the vowel. Since this is an open CV syllable, the movement cannot be attributed to any transition to a final consonant.

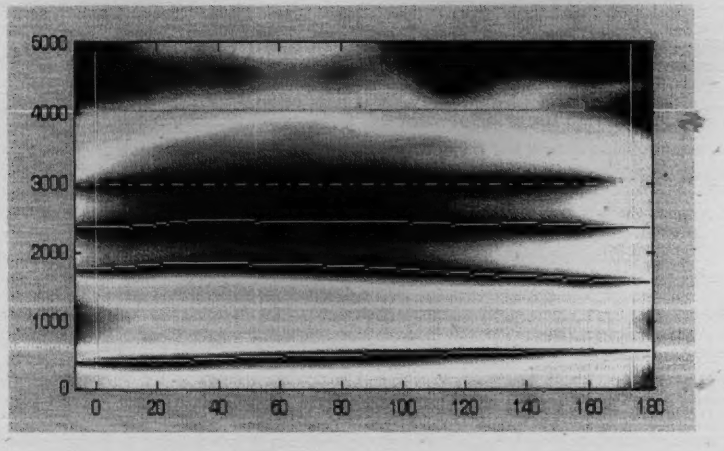


Figure 2.3. A female native English speaker's [ɪ] production extracted from [bɪ] syllable.

In Figure 2.4, average F1 and F2 values at both the beginning and end of English vowels and their trajectories are plotted. It is clear from this illustration that most English vowels are not true monophthongs. Apart from the importance of incorporating VISC for F1 and F2 into a model of crosslinguistic vowel similarity, other dimensions such as F3 and vowel duration may also have an effect on the perceived degree of crosslinguistic similarity. However, these variables are not incorporated in the two-dimensional models most often used in research within the SLM paradigm.

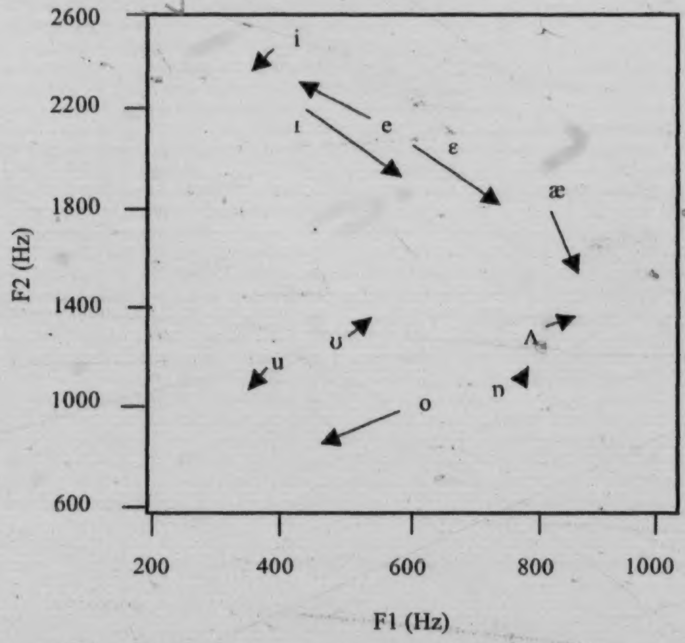


Figure 2.4. F1 and F2 formant values from the beginnings and endings of Canadian English vowels as well as their trajectory indicated by the arrows, borrowed from Nearey and Assmann (1986). Reproduced with permission from first author.

In addition to the failure to represent phonetic multidimensionality, there are other important limitations to this commonly used approach for determining crosslinguistic similarity. While the SLM explicitly recognizes that all production tokens in an L2 are not equally good members of the learner's L1 categories, in practice, SLM research often treats all instances of a given L2 category as equally representative of that category on the basis of elliptical distribution boundaries. For example, in Bohn and Flege's (1992) study, they drew ellipses around two-dimensional distributions for English /i/ and German /i/, and then concluded from the degree of overlap in the distributions that German speakers would perceive productions of English /i/ to be good members of the German /i/ category. In fact in Bohn and Flege (1992), the elliptical boundaries around English /i/ clearly did not fall within the German /i/ norms. Even more confusing was the overlap between English /ɪ/ and German /ɛ/ and /ɛː/. It appears from Bohn and Flege's comparison that /ɛ/ and /ɛː/ could just as easily be described as 'similar' to English /ɪ/ as to English /ɛ/. Yet, the researchers conclude that German /ɛ/, /ɛː/ are 'similar' to English /ɛ/, not /ɪ/. This certainty appears to be derived from a priori assumptions about which categories are most likely to interact. This raises another limitation of much previous research within the SLM paradigm: L1/L2 comparisons are often made on the basis of a limited subset of categories. This selective analysis may lead to bias, or failure to recognize potential interactions with other L1 and L2 categories – interactions that might be discovered if comparisons are made across larger portions of the vowel space.

Finally, the acoustic comparisons of vowel inventories commonly used also fail to incorporate cross-speaker variation stemming from differences in the size and shape of different speaker's vocal tracts. Strange (2007) points out the potential hazard in ignoring the speaker normalization problem. It is possible that researchers may find differences in spectral properties across languages that are not related to category differences, but rather, to differences in the speakers who produced them. Although part of this speaker normalization problem may be resolved with adequately large sample sizes, a better approach would be to incorporate speaker variables into measures of crosslinguistic similarity.

Morrison (2006) applied discriminant analysis to measure similarity between Spanish and English vowels. Morrison's study incorporated the sort of multidimensional acoustic cues I am arguing are often critical for vowel identification. First, using acoustic measurements from L1 English vowel productions (using measures of F1, F2 and F3 from the beginnings and ends of vowels, duration, and F0 to normalize for speaker variability), he built a statistical pattern recognition model which defined English vowel categories and then tested Spanish vowel productions against this English model to determine how the model would classify each production. He reversed this process to determine how English vowels would be classified by a Spanish statistical pattern recognition model. His results provide an indication of how Spanish and English learners might perceive the opposing language's vowels. Strange (2004, 2007) has applied a similar approach to classifying French and German vowel productions in terms of American English categories with mixed results. My own first attempt at applying this procedure to English and Mandarin vowels is reported in Thomson (2005). A limitation of this approach is that while it provides an indication of the closest vowel categories across languages, it does not indicate how well a production of a vowel in one language fits a vowel category in another language. It also assumes that an L2 learner will be forced to classify all L2 vowel productions in terms of an L1 category, which may not be the case. It is possible that learners may recognize that some L2 vowels are simply not sufficiently adequate examples of any L1 vowel category. If learners are able to recognize some L2 vowels as deviant from all L1 categories, they may be encouraged to attempt to form a 'new' L2 category that includes such deviant tokens.

2.3. A statistical pattern recognition approach to L2 phonological learning

2.3.1. Measuring crosslinguistic similarity

In this section, I propose a further extension of the statistical pattern recognition approach to crosslinguistic similarity, one that is more detailed than most previous approaches that rely on acoustic measures for ascertaining degree of crosslinguistic similarity. This approach allows for analysis of similarity across a larger number of dimensions than is achieved on the basis of the F1/F2 plots most commonly used in L2 speech research. In addition, it enables comparison of *all* vowel categories across the L1 and L2 rather than relying on a priori assumptions concerning which L2 categories are 'similar' or 'new'. As I discussed in the preceding section, while I continue to frame this discussion in terms of crisp 'same', 'similar' and 'new' distinctions, I do so for practical reasons only. These terms do not reflect my ontological orientation.

A pattern recognition approach to vowel categorization has been effectively used to describe L1 English vowel perception in a number of studies (Assmann, Nearey, & Hogan, 1982; Assmann & Katz, 2000; Hillenbrand & Nearey, 1999; Nearey & Assmann, 1986). Illustrated in Figure 2.5, it relies on dynamic information from the vowel: measures of F1 and F2 and F3 taken from points at both the beginning and end of the vowels, as well as F0 (pitch) and vowel duration. Dynamic information is used because, as mentioned previously, it has been determined that diphthongization, at least for English, serves as an important cue to vowel identification (Assmann et al., 1982). F0 is used as a means to account for speaker differences – particularly gender. Using discriminant analysis, the observed values from relevant variables (F1, F2, F3, pitch and duration) in 'training' data are used to model predicted group membership of 'new' cases.

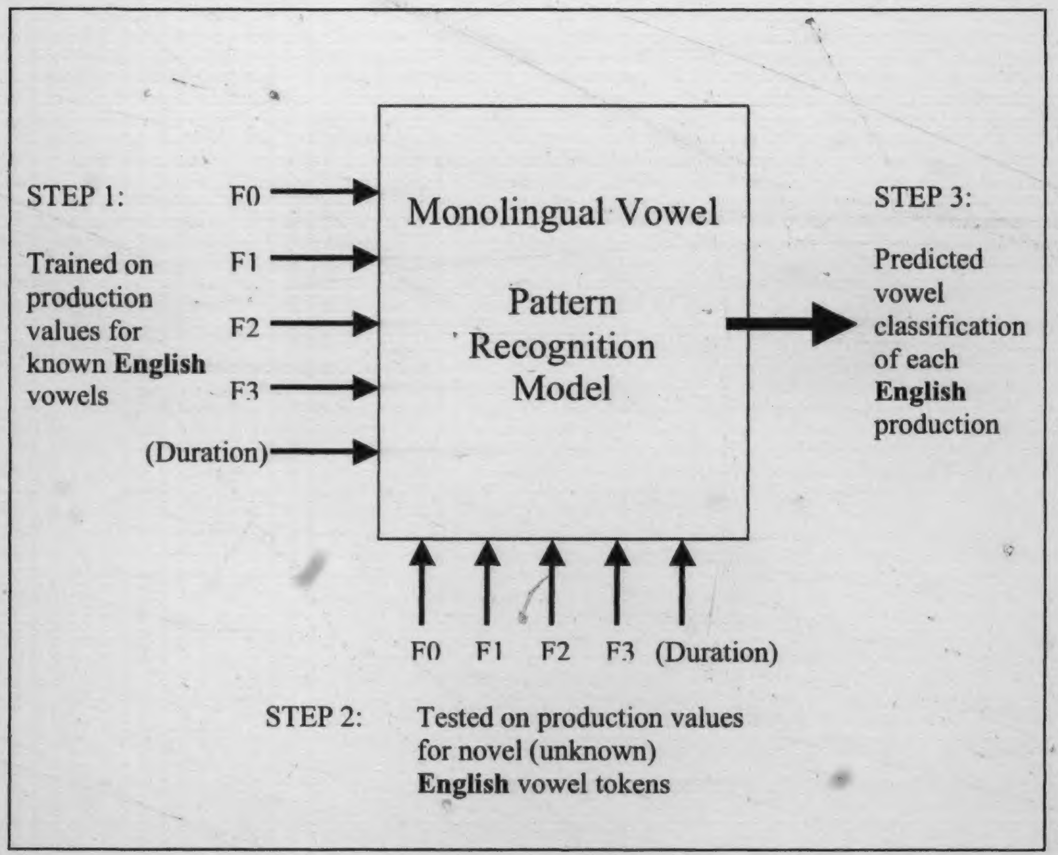

STEP 1:
Trained on production values for known **English** vowels
F0
F1
F2
F3
(Duration)
Monolingual Vowel
Pattern Recognition Model
STEP 3:
Predicted vowel classification of each **English** production
F0
F1
F2
F3
(Duration)
STEP 2:
Tested on production values for novel (unknown) **English** vowel tokens

Figure 2.5. Monolingual vowel pattern recognition model

The predicted categorizations of new cases by L1 pattern recognition models have been previously shown to be highly correlated with vowel stimuli categorizations by human listeners (Assmann et al., 1982; Hillenbrand & Nearey, 1999; Nearey & Assmann, 1986). For example, the product moment correlation between human listeners and pattern recognition classifications in Nearey and Assmann (1986) was r = .990 for natural, full vowels.

In effect, an L1 pattern recognition model of this type can be viewed as a sort of idealized perceptual model – what a listener may perceive under ideal circumstances with no biases from other sources. Crucially, previous research applying this pattern recognition model to L1 vowels has demonstrated that this statistical approach offers a parametric representation of English vowels that reasonably approximates human vowel perception. It may not account for a human listener's response to speech in noisy conditions, however, or a lexically or contextually motivated bias to categorize an otherwise ambiguous production in a particular direction.[2]

This type of pattern recognition model can be applied to the issue of measuring L1/L2 vowel similarity by creating a crosslinguistic model, henceforth termed the Metamodel. The Metamodel incorporates categories from both the L1 and the L2. This is illustrated in Figure 2.6 below. The extent to which the Metamodel should be interpreted as having ontological status vis-à-vis L2 phonological systems will be touched upon shortly. First, I will sketch out how the model may be applied to more precisely measure statistical differences between L1 and L2 sounds in the spectral dimension.

---

[2] As mentioned earlier, Strange (2004, 2007) has applied discriminant analysis using acoustic measurements to assess L2 vowel productions and found that while a relationship to human listener perceptual responses exists, the relationship is not reliable in all cases. On the surface, this appears to be a contradiction to the claims of Nearey and colleagues that a discriminant analysis pattern recognition approach is strongly correlated with human listener responses. It should be made clear that Strange's (2004, 2007) approach is not the same as that used by Nearey and colleagues. Most importantly, Strange relies on static formant frequency measures taken from the midpoint of vowels rather than dynamic information that has been shown to be crucial to accurate vowel identification by human listeners, at least for English vowels (Assmann et al., 1982). Consequently, Strange's findings may not support her claim that acoustic measurements do not accurately reflect human responses (Strange, 2007). Rather, discrepancies between her acoustic measurements and human responses seem just as likely to result from a reliance on incomplete acoustic information that she has used to build her statistical model. Although comparing acoustic measures with human listener responses is beyond the scope of this dissertation, earlier work comparing the human listener responses to data in Munro et al. (2003) with acoustic measures in Thomson (2005) suggests that with dynamic acoustic information, a discriminant analysis model of the sort employed by Nearey and colleagues does indeed correlate with human listener responses to L2 data.

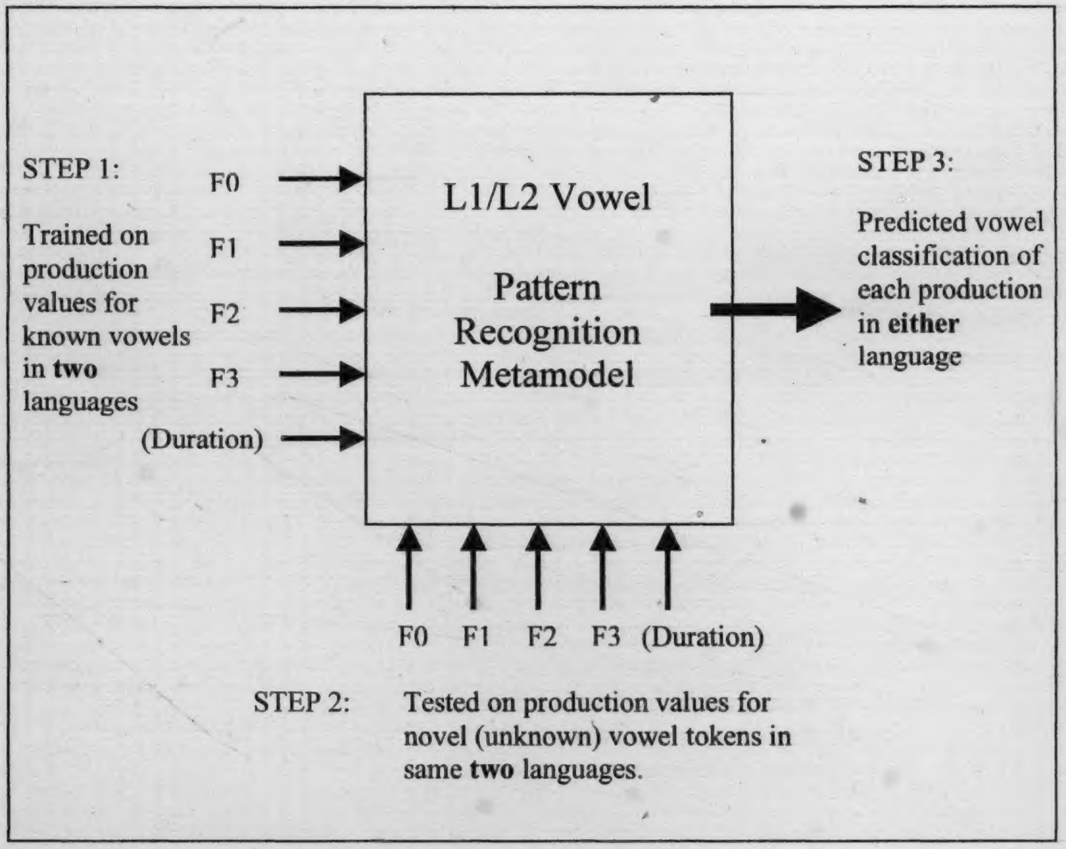

STEP 1:
Trained on production values for known vowels in **two** languages
F0
F1
F2
F3
(Duration)
L1/L2 Vowel
Pattern Recognition Metamodel
STEP 3:
Predicted vowel classification of each production in **either** language
F0
F1
F2
F3
(Duration)
STEP 2:
Tested on production values for novel (unknown) vowel tokens in same **two** languages.

Figure 2.6. Two-language vowel pattern recognition Metamodel.

After training the Metamodel on all relevant categories from both languages, production values for new cases from each language can then be tested. The extent to which new cases from one language are misclassified as members of a category in the opposing language provides a means of determining crosslinguistic similarity. For example, if the general distribution of a phonological category in one language were truly *identical* to the distribution of a category in the competing language, we would expect that in the Metamodel, 50% of new cases of the category in the first language should be misclassified as members of the corresponding category in the competing language and vice versa. If the general distribution of two categories is not identical, but very similar and results in the predicted merger of the L1 and L2 categories, we should expect to find that L1 and L2 productions of the hypothesized category have the same distribution across Metamodel categories regardless of the language being spoken. Furthermore, the distribution of the merged category should not be the same as that of monolinguals' productions of the merely 'similar' categories in each language.

In addition to crisp, absolute classification of 'new' cases, statistics used in the model can tell us something about how well a 'new' case fits the category. In discriminant analysis, the assignment of a 'new' case to a specific category is determined on the basis of linear classification functions that establish the *a posteriori probability* (APP) of each new case's membership in a specific category. For example, while a new case may be classified as category X by the model, the APP value of that case's membership in that group may be only .51, while the probability of membership in competing category Y, or the sum of all competing categories, is .49. It can be assumed that such a case is a relatively poor example of the category to which it has been classified, relative to a case with an APP of group membership of .95. APPs, then, provide a way of determining how well a particular production fits the category to which it is assigned vis-à-vis competing categories. For the purposes of comparing two vowel systems, we would expect 'similar' vowel tokens that are accurately classified as the intended language category in absolute terms to also have some probability of being members of the 'similar' category in the opposing language. More specifically, 'similar' but not identical vowel tokens will have higher APPs of membership in an opposing language category, while dissimilar vowel tokens will have lower APPs of membership in any opposing language category (see Nearey & Assmann, 1986 for a more detailed description of APP scores).

As well as providing a measure of crosslinguistic similarity, a pattern recognition model also has the ability to estimate the degree to which accented productions approach L2 targets. Studies of L2 phonological development, or accentedness, have largely relied on intelligibility scores obtained from native speaker listeners. While this approach has important strengths in that results are indisputably representative of NS responses to accent, there are also limitations in the hypotheses that can be tested. Even trained human listeners are likely affected to some degree by the same perceptual assimilation processes evident in L2 learning. This means that their perception of L2 accented speech will often be categorically determined. It is relatively easy for a native speaker listener to assign an accented production to a particular target language category, as was demonstrated in Munro, Derwing and Thomson (2003). It is more difficult, however, to determine to what degree the accented production fits into the target category, particularly the subtle changes associated with movement over time from less nativelike to more nativelike. Testing accented productions against a Metamodel that incorporates the speakers' L1 as well as L2 categories, we may be able to gain insight into the extent to which L2 learners are simply producing L1 sounds in the L2 versus the extent to which they are producing sounds that more likely reflect the development of new L2 categories. In addition, APPs provide a way to measure how close L2 productions are to the target category.

The same variables used in discriminant analysis can also be used to derive Mahalanobis Distance (MD) scores[3]. These measures can provide additional information, in absolute terms, concerning how far from target category centres each 'new' case is. As such, they provide an additional 'goodness of fit' measure for accented productions. For example, whether an accented production is correctly classified as its intended target or not, a Mahalanobis Distance score provides information about how far it is in absolute terms from that categories centre and therefore can be used to more precisely measure improvement in accent over time without having to rely on the absolute classification of a production as either 'correct' or 'incorrect' that depend entirely on previously established category boundaries. Instead, whether the phonetic properties of a given L2 production fall within the distribution of the intended category or not, MD scores specify the production's precise distance from the category's centre.

---

[3] Mahalanobis Distance is defined as: $D^2 = (\bar{X}_1 - \bar{X}_2)'S^{-1}(\bar{X}_1 - \bar{X}_2)$ where $\bar{X}_1$ and $\bar{X}_2$ are the mean vectors of two groups being compared. S is the weighted average of the variance-covariance matrices for the two groups.

2.3.2. The Metamodel and L2 phonological representations

The Metamodel is clearly not representative of real-world perceivers in the way an L1 statistical pattern recognition model is. Rather, it represents what an idealized bilingual human listener would be able to perceive in a perfect interlanguage, where small phonetic differences between similar L1 and L2 categories were actually discernable. It may be conceived of as the ideal end-state of any bilingual whose phonological system allowed him/her to develop distinct L1 and L2 categories.

In reality, this sort of theoretically ideal bilingual end-state has not been attested. Flege's (1995) SLM has so far been proven correct in its claim that when a category in a learner's L1 is very similar to a category in the learner's L2, establishing the L2 category as a separate category is difficult. Instead, it is more likely the case that in such a scenario, a learner will incorporate the L2 category into his/her L1 category, resulting in a single category that is used in both languages, with values that are intermediate between the L1 and L2 distributions. In the Metamodel, such merged categories would result from instances where the statistical distance between an L2 category and a pre-existing L1 category is insufficiently large to allow the learner to perceive the L1 and L2 categories as meaningfully different. The precise details of such L1/L2 mergers are beyond the scope of this dissertation. In general, however, it should be understood that when the Metamodel provides evidence of substantial overlap or confusion between an L1 and an L2 category, such categories are very likely to undergo a merger in most learners' phonological systems rather than developing as two distinct categories as the Metamodel may imply.

In contrast to 'similar' categories, categories in the L2 that are 'new' or less 'similar' to existing L1 categories should eventually emerge as distinct categories in the learner's phonological system. In such cases, the Metamodel represents these categories as L2 categories whose statistical distance from pre-existing L1 categories is sufficiently large that they are rarely if ever confused with any L1 category. Such statistically recognizable differences between L1 and L2 categories may reflect a discernable difference available to L2 learners. If discernable, such differences may provide learners with the motivation for developing a 'new' L2 category rather than substituting an L1 category in place of the ‘new’ category, as they often do with ‘similar’ categories.

2.4. Summary

In this chapter, I have argued that our ability to assess the effect of a learner’s L1 on L2 phonological learning is limited by the extent to which we can adequately operationalize crosslinguistic similarity. Having proposed a statistical pattern recognition approach that I believe provides a better way of assessing degree of crosslinguistic similarity, in the next chapter, this approach will be applied to a comparison of English and Mandarin vowels. The results of this comparison will be used in later chapters to test specific predictions of learner behavior in the context of a study of Mandarin learners of English vowels.

Chapter 3. A comparison of English and Mandarin vowel systems

Mandarin Chinese and English vowel systems are ideal for the purposes of this study because not only are Mandarin and English vowel inventories very divergent, these languages are very different across other linguistic domains (e.g., syntax, morphology, lexis). Much of the previous research examines phonological learning in related languages. Piske, MacKay and Flege (2001) summarize a number of published studies of L2 accent on the basis of the learners' L1s. Most of the studies listed investigate L2 learners from related L1 groups (e.g., English, Spanish, Italian, Dutch, French and German) with far fewer devoted to speakers of unrelated L1s (e.g., English, Arabic, Japanese, Persian, Thai, Russian, Korean and Mandarin). The bias toward studying related languages limits our understanding and our ability to make generalizations across L2 learners. If languages are related in terms of syntax, morphology and lexis for example, learners may be able to direct greater attention to areas of divergence, such as phonology.

Recall that on the basis of a study of adult Dutch learners of English discussed earlier, Bongaerts et al. (1997) concluded that given ideal learning conditions, adult L2 learners are capable of achieving nativelike pronunciation in an L2. Bongaerts et al.'s (1997) conclusions regarding adult L2 learners' ultimate ability would be more convincing if they could demonstrate that adults L2 learners from a typologically distant L1 (e.g., Vietnamese or Zulu) are also capable of achieving a nativelike accent in an English L2. In fact, in a follow-up study, Bongaerts, Mennen and van der Slik (2000) examine this possibility and arrive at slightly weaker conclusions regarding adult L2 learners' ability to develop a nativelike accent. In this later research, Bongaerts et al. (2000) examined the pronunciation of thirty advanced proficiency adult learners of Dutch from a variety of L1 groups. They found that given sufficient exposure to Dutch, many learners achieved a nativelike accent. However, they also found an apparent relationship between a learner's L1 and his or her ultimate pronunciation ability in the L2. The eleven Dutch learners in Bongaerts et al.'s study who attained near nativelike ability, with the exception of one Czech participant, spoke L1s that were closely related to Dutch (i.e., German, English and French). Apart from the Czech participant, no other learners from a less similar L1 group (i.e., Armenian, Berber, Greek and Turkish) were able to speak Dutch without a detectable accent. Consequently, Bongaerts and his colleagues concluded that typological similarity may be a determining factor in ultimate attainment.[4] Additionally, many of the learners in Bongaerts et al. (2000) had limited if any formal instruction in Dutch. This led the researchers to suggest that instruction may also play a facilitative role. In other words, even when massive in quantity, naturalistic input may not be sufficient for the development of accent-free L2 speech production.

The choice of Mandarin learners of English for my study is intended to contribute to knowledge of L2 phonological learning by speakers of L1s that are more distinct from the L2. Research measuring the ability of Mandarin learners of English to discriminate English vowel contrasts found that even among those adult learners who had been in the United States for between three and five years, ability in production was far from nativelike (Jia, Strange, Wu, Collado & Guan, 2006). Therefore, if my study results in improvement in perception and production of English vowels by Mandarin speakers, this might indicate that the apparent difficulty Mandarin speakers face in reaching more nativelike English targets can be mitigated through instructional intervention. Finally, my study builds upon earlier research of Mandarin learners of English in which I have been involved (Munro et al. 2003; Munro & Derwing, 2007; Thomson, 2005).

Another motivation for this study is a desire to extend the scope of L2 speech analysis to a larger set of phonological categories. As mentioned in the last chapter, most previous research has been limited to comparing only small subsets of L1/L2 phonological systems; the scope of research is usually limited to contrasts that have been known for some time to cause difficulty for learners. For example, in the study by Bohn and Flege (1992) discussed earlier, English /i/, /ɪ/ and /ɛ/ were contrasted with similar German categories to determine the nature of L1-based difficulties for German learners of English. In a previous study of Mandarin and Cantonese L1 learners of English (Wang & Munro, 2004), it was shown that Mandarin learners have difficulty with specific

---

[4] It is unclear from Bongaerts et al. (2000) how exactly the issue of typological similarity impacts L2 phonological acquisition. For example, the vowel systems of some typologically similar L1s (e.g., English and Spanish) are arguably as dissimilar from each other as are the vowel systems of typologically distant L1s (e.g., English and Mandarin). It may be that when two L1s are typologically similar, there is generally more potential for positive transfer in other linguistic domains and this may facilitate greater attention by learners to domains that are more distinct.

English vowel contrasts, namely, /i/-/ɪ/, /u/-/ʊ/ and /ɛ/-/æ/. My study expands these contrasts to examine a much larger portion of the English vowel space while making fewer a priori assumptions about how specific categories in the L1 interact with categories in the L2. This means it has the potential to discover L1/L2 interactions that may exist beyond those that are immediately apparent.

The aim of this experiment is two-fold. First, application of the statistical pattern recognition model proposed in the previous chapter will provide a basis for predictions concerning which vowels in English will be most difficult for Mandarin speakers to acquire in a nativelike fashion. In terms of overall identification patterns, Best's (1995) PAM predicts that those English categories that directly assimilate to a single Mandarin category will be most accurately identified by learners in an identification test. If two English categories assimilate to a single Mandarin category, those English categories will be less accurately identified by Mandarin learners. If an English category has no obvious counterpart in Mandarin, it may or may not be accurately identified, depending on, among other factors, its interaction with other English categories. In particular, errors in identification between English categories may result for English tokens that are somewhat ambiguous with regard to two English categories (e.g., /ɪ/ vs. /ɛ/). Although PAM does not specifically address this possibility, it is widely accepted that limited within-language ambiguity can exist for some phonemic contrasts (e.g., Bond, 1999). In terms of ultimate learnability, following Flege's (1995) SLM, those vowel categories that have the largest number of shared acoustic properties will be defined as most similar and are hypothesized to present the greatest challenge in the long term, although vowels that are least similar may also present a challenge in the short term.

While PAM and SLM will guide my analysis, they are used as general frameworks – a way to contain discussion. However, when either is mentioned, they should be reinterpreted in terms of my statistically defined similarity continuum, where individual production tokens of the same nominal category can show graded behaviour in terms of how the Metamodel and listeners respond to them. Applying the Metamodel approach to Mandarin-accented English productions for example, specific L2 English productions of categories that are 'similar' to Mandarin categories will more often be recognized by the Metamodel as members of the 'similar' Mandarin category than as productions of the intended English category. This is due to the learners' predicted substitution of L1 sounds in the production of 'similar' L2 categories. In contrast, L2 English productions of categories that tend not to be very similar to Mandarin categories, although sometimes inaccurate, are more likely to be recognized by the Metamodel as other English categories, rather than as members of Mandarin categories. This is because less 'similar' or 'new' categories will begin to emerge that are distinct from any Mandarin category. These new categories are represented by the Metamodel as those English categories that have little if any confusion with any Mandarin categories. These predictions will be specified in much greater detail in chapter 5.

The second aim of this experiment is to establish a ranked set of natural English vowel tokens for each English vowel category in terms of their degree of similarity to *any* competing Mandarin vowel category. In naturally varying productions, some tokens of a vowel category are easier to perceive than others. If these differences stem from the degree to which an individual production token is similar to a competing L1 Mandarin category, then we might expect that Mandarin L2 English learners should be better able to identify those tokens in production that are furthest from *any* competing Mandarin category because they are more likely to notice that they are 'new' or otherwise different. Thus for example, production tokens that are furthest from any competing Mandarin category might be most likely to escape from any magnet effects such as those proposed by Kuhl and colleagues (Kuhl, 2004; Kuhl & Iverson, 1995). In terms of the Metamodel, such English production tokens will not only be correctly classified as the intended English category vis-à-vis competing Mandarin categories, but they will also have relatively low APPs of being a competing Mandarin category. These predictions will be tested in a training experiment described in later chapters.

3.1. Method

3.1.1. Mandarin and English vowel inventory selection

It is maintained in the SLM that perceptual errors (and resulting errors in production) are caused by interference from L1 position-sensitive allophones (Flege, 1995); that is, the perception of a category is dependent on the specific context in which the phoneme is found. For example, following this reasoning, the SLM predicts that perception of a vowel in a post-obstruent context will not necessarily extend to perception of the same vowel in a post-fricative context. This seems to contradict vowel perception research which indicates that while dynamic information found in consonant-vowel coarticulation is important, this information is invariant across consonantal contexts (Bohn and Polka, 2001; Jenkins, Strange & Trent, 1999; Strange, 1989). Even the effect of consonant is disputed; Andruski and Nearey (1992) found that for English speakers, vowel transitions from and to consonants were not of great importance. Rather, they claim that vowel inherent spectral change is usually sufficient for identifying English vowels.

There is evidence that spectral properties of vowels vary across consonantal contexts. Strange (2007), for example, demonstrated that differences exist between productions of the same English vowels produced in [hVbə], [gəbVpə] and [gədVtɔ] contexts. While native speakers are able to recognize similarities between the same vowels produced in different contexts, it is unclear to what extent beginning proficiency L2 learners are able to adjust for such contextual effects. Rochet (1995) found that Mandarin speakers' learning of a French word initial stop voicing contrast transferred from a bilabial context to alveolar and velar contexts. However, while transfer of this voicing contrast occurred in word initial position, it did not appear to transfer to word medial position. Conversely, Broersma (2005) found that the ability to perceive a voicing contrast in Dutch word-initial position transferred to word final position for Dutch learners of English, despite the fact that Dutch lacks a voicing contrast in word final position.

Despite conflicting evidence regarding SLM's claim that phonological learning takes place in relation to positionally-sensitive allophones, I have decided to take a conservative approach and limit my vowel selection to those that occur in very similar Mandarin and English contexts. This provides a modicum of control. Two minor exceptions are made for reasons discussed below.

The Mandarin and English vowel categories chosen for comparison were determined by identifying all Mandarin and English vowels found in the context of bilabial consonants (i.e., bV and pV sequences). This particular CV context was chosen because it afforded a modicum of control for crosslinguistic similarity. Mandarin /p/ is described phonetically as an aspirated voiceless bilabial, while Mandarin /b/ is described phonetically as an unaspirated voiceless bilabial (Li and Thompson, 1997). While English word initial /b/ is sometimes pronounced with a negative VOT, I assume for the purposes of this study that the difference with Mandarin is negligible in terms of its effect on the vowel. Open CV syllables were chosen because Mandarin has no coda consonants apart from nasals and possibly glides. Using similar syllable structure is important since the ultimate goal is to predict Mandarin L1 transfer to English L2.

Given these explicit constraints, choosing relevant Mandarin vowel categories for analysis was still not entirely straightforward. According to Lee and Zee (2003), Standard Mandarin contains a six-vowel inventory that in IPA terms are represented as /i, y, ɔ, a, ɤ, u/. Other sources describe the Mandarin inventory as including up to eight vowels: /i, y, e, ə, a, o, ɤ, u/ (Chen, 1976; Maddieson, 1984). Part of the discrepancy may stem from decisions concerning what are monophthongal vowels and what are diphthongs. Lee and Zee's (2003) list of diphthongs includes /ou/ and /ei/, two categories that are listed as monophthongal vowels (i.e., /o/ and /e/) by Maddieson (1984) and Chen (1976). In addition, wide-ranging allophonic variation in Mandarin complicates the choice of sounds for comparison to English counterparts. Referring to further descriptions by Duanmu (2003) and in consultation with a native speaker of Mandarin with expertise in phonetics (Zhang, personal communication, February 2006), I determined that only five Mandarin vowels listed by Chen (1976) and Maddieson (1984) were found in pV or bV contexts: /i, e, a, ɤ, u/. In addition, I found that although they did not occur in pV or bV contexts, two other vowels, /o/ and /uɔ/,[5] were also of potential interest for comparison to English vowels. Although these last two do not occur in post-labial contexts, they do occur after alveolar and velar obstruents. Consequently, I felt it was reasonable to include them in the crosslinguistic similarity model because the context in which they are found is related in manner, and therefore a possibility exists that they might aid in the learning of similar English vowels.

---

[5] Mandarin /uɔ/ listed by Lee and Zee (2003) as a diphthong is not the same as the Mandarin monothong /u/ which they also list and which was also selected for analysis.

While it might also seem reasonable to consider Mandarin /y/ as being potentially similar to English /u/ or /i/, in this study Mandarin /y/ was not included in the analysis. My decision to exclude it was based on earlier research in which Mandarin /y/ was included. I found that this vowel was not statistically similar to English /u/ and it was only minimally similar to English /i/ (Thomson, 2005). Furthermore, in that research, L2 productions of English /u/ were never confused with Mandarin /y/. One possible explanation for the lack of interaction between Mandarin /y/ and English vowels is that this Mandarin vowel has a very limited distribution in Mandarin and is never found in /bV/ or /pV/ contexts, which was the focus of my earlier research; Mandarin /y/ is also not found in /dV/, /tV/, /gV/ or /kV/ contexts.

After selecting the seven Mandarin vowels /i, e, a, uə, o, ɤ, u/ for comparison to English, it was further determined that the Mandarin data should be produced in the 4th tone which has a high falling tone contour. This decision was made after placing pilot recordings of the target syllables in the English carrier, “The next word is ____”, where the stress is placed on the target word in final position. Two native speakers of English with phonetic expertise determined that productions in the fourth tone of Mandarin best fit this English stress pattern. This was deemed important since I planned to elicit the English data in the same English carrier phrase, which has phrase final falling intonation.

The English vowel categories chosen were /i, ɪ, e, ɛ, æ, ɒ, ʌ, ʊ, o, u/ the same ten Albertan English vowels identified by Nearey and Assmann (1986). Apart from the exclusion of /ɔ/ as a category, the rest are identical to Hillenbrand, Getty, Clark and Wheeler’s (1995) description of American vowels.

Finally, it should be noted that 10 of the resulting 14 Mandarin syllables (7 vowels in both /bV/ and /pV/ frames) were real words in Tone 4; the same 10 Mandarin syllables are the only real words in the remaining tones as well. Only 8 of the 20 English syllables are real words. More details follow in the procedure section.

3.1.2. Speakers

Mandarin L1 speakers

The Mandarin vowel production data were obtained from 20 native Mandarin speakers (10 male, 10 female; ages 20-46, *M* = 28.2) who were from Mainland China and who all reported speaking a standard variety of Mandarin. All were current or former students at the University of Alberta. For admission to the University of Alberta, a minimum TOEFL score of 580 on the paper-based test, 237 on the computer-based test or 86 on the internet based test (with a score of at least 21 on each band) is required; this indicates that these Mandarin speakers were all high proficiency English speakers. All those chosen for participation reported speaking a standard dialect of Mandarin. Their length of residence in Canada ranged from three months to six years (*M* = 2.55 years) and their age of arrival between 18 and 44 years of age (*M* = 25.65 years). Although all had completed some post-secondary studies in English, only ten had ever taken official ESL classes since arriving in Canada (range 0 to 1 year, *M* = 6 months). All reported normal hearing.

Native English speakers

The English vowel production data were obtained from 20 native English speakers from the undergraduate student population at the University of Alberta (10 male, 10 female; ages 18-50, *M* = 28.55). All had resided in Western Canada (most in Alberta) since their childhood and had spent the majority of their lives there. In addition, while several reported advanced knowledge of a second language, they all used English as their primary language. All reported normal hearing.

3.1.3. Procedure

Mandarin productions

Participants from the Mandarin native speaker group were recorded individually in a quiet room, using a high quality Marantz digital recorder with a sampling rate of 41,110 Hz. Participants were asked to listen to and repeat a series of /bV/ and /pV/ stimuli containing the target vowels spoken by a female native speaker of a standard variety of Mandarin. All stimuli were produced in the fourth tone of Mandarin. These targets are shown in Table 3.1.

All syllables, except those denoted with an asterisk, are real Mandarin words. Those denoted with an asterisk are not Mandarin words, nor are they possible Mandarin syllables. As mentioned earlier, these target vowels never occur after a labial obstruent in any Mandarin word. To insure that participants understood that they were to put a Mandarin vowel in this nonsense syllable, immediately prior to their hearing the target prompt, I provided them with real word prompts that rhymed with the target nonsense syllable. The real word prompts with alveolar and velar onsets were: dè, gè, tè, kè, dòu, gòu, tòu, and kòu.

Table 3.1. Syllables used for elicited imitation of seven target Mandarin vowels.

	Mandarin /bV/ targets						
IPA	bi	be	ba	buə	bo	bɤ	bu
Pinyin	bì	bè	bà	bò (buò)	bòu*	bè*	bù
	Mandarin /pV/ targets						
IPA	pi	pe	pa	puə	po	pɤ	pu
Pinyin	pì	pè	pà	pò (può)	pòu*	pè*	pù

The Mandarin speakers demonstrated no difficulty understanding or carrying out this task. Their productions of Mandarin /o/ and /uə/ in the target labial context mirrored their productions in the Mandarin real-word contexts. This provided confirmation that although Mandarin /o/ and /uə/ do not exist in the target postlabial context, speakers have no difficulty transferring them from related contexts. The speakers' ability to successfully produce phonotactically illegal Mandarin syllables using Mandarin phonemes provides further justification for expecting that Mandarin L2 English learners should also be able to transfer this knowledge from Mandarin postalveolar and velar obstruent contexts to English postbilabial obstruent contexts.

All stimuli were presented in a Mandarin carrier phrase "Xia yige zi shi _____[6]" that, translated, means "*The next word is* _____" and participants were asked to respond by repeating the word in a Mandarin carrier, “Xianzi wo shuo ____” that, translated, means, “*Now I say* _____”. In addition to the auditory prompt providing a pronunciation model, each word was also provided in written form in Pinyin (and Chinese characters when possible). The entire procedure was repeated twice for each participant in order to record two repetitions of each item. After recording each of the speakers’ productions, the target syllables were extracted from the sentence frame. Next, they were down-sampled to 22.055 kHz, normalized across tokens, and saved as a separate sound files for each syllable.

---

[6] As transcribed in Pinyin.

English L1 productions

A similar procedure to that outlined above for Mandarin was used to gather production data for each of the ten English vowels being analyzed. Participants from the English native speaker group were tested individually in a quiet room, and recorded using a high quality Marantz digital recorder with a sampling rate of 44,110 Hz.

Participants were asked to listen to and repeat a series of /bV/ and /pV/ stimuli containing the target vowels spoken by a female native speaker from Edmonton, Alberta. These targets are shown in Table 3.2.

Table 3.2. Syllables used for elicited imitation of ten target English vowels.

	English /bV/ targets									
IPA	bi	bɪ	be	bɛ	bæ	bɒ	bʌ	bo	bʊ	bu
	English /pV/ targets									
IPA	pi	pɪ	pe	pɛ	pæ	pɒ	pʌ	po	pʊ	pu

Although eight of these English syllables are real words (i.e., be, bay, bow, boo, pea, pay, paw, poo) the majority are nonce words. To minimize the number of potential errors, each participant was asked to pay particular attention to the vowel portion. It should be noted that although open /bV/ and /pV/ syllables containing lax vowels violate English phonotactic constraints, for the most part speakers’ had little difficulty completing this task accurately.

All stimuli were presented in the carrier phrase, "*The next word is* _____" and participants were asked to respond by repeating the word in the carrier, "*Now I say* _____". Since there is not a straightforward way to represent English nonce words to participants who are not familiar with a phonetic alphabet, it was not possible to provide participants with a written form of the target syllables. The entire procedure was repeated twice in order to record two repetitions of each item. After recording each of the speaker's productions, the target syllables were extracted from the sentence frame. Next, they were down-sampled to 22.055 kHz, normalized across tokens at 50% peak amplitude, and saved as separate sound files for each syllable.

The vowel elicitation procedure used in this study is different from the common use of production elicitation using orthographic or phonetic alphabet prompts (cf. Hillenbrand, Getty, Clark & Wheeler, 1995) with no accompanying auditory prompt. I chose to present the stimuli auditorily to provide a model. Having the stimuli presented in a sentence frame with the response also requiring a sentence frame made it more likely that this task would result in processing and reproduction of the stimuli in each native speaker's own phonological systems rather than through simple mimicry.

3.1.4. Data analysis

The first repetitions of each item for each language were screened to insure recording quality was satisfactory and vowel quality was of the intended target. Only one token from the 280 first repetitions (20 speakers x 14 target syllables) of the Mandarin recordings was replaced with the second repetition by the same speaker, a production of /uə/ that sounded closer to Mandarin /ɤ/. For the English tokens, seven of the 400 first repetitions (20 speakers x 20 target syllables) were replaced with second productions, four /ɛ/ productions and three /ʌ/ productions. In these cases, the first repetition was unsatisfactory and the second repetition was only slightly better. It should be mentioned that based on my own perceptual screening, many more English productions were not as target-like as I had anticipated, particularly for these two English vowels, /ɛ/ and /ʌ/.

Using a suite of acoustic analysis tools created with Matlab by T. M. Nearey (with modifications by Geoff Morrison), vowel boundaries were marked for each sound file (400 for English and 280 for Mandarin). This procedure is visually illustrated in Figure 3.1.

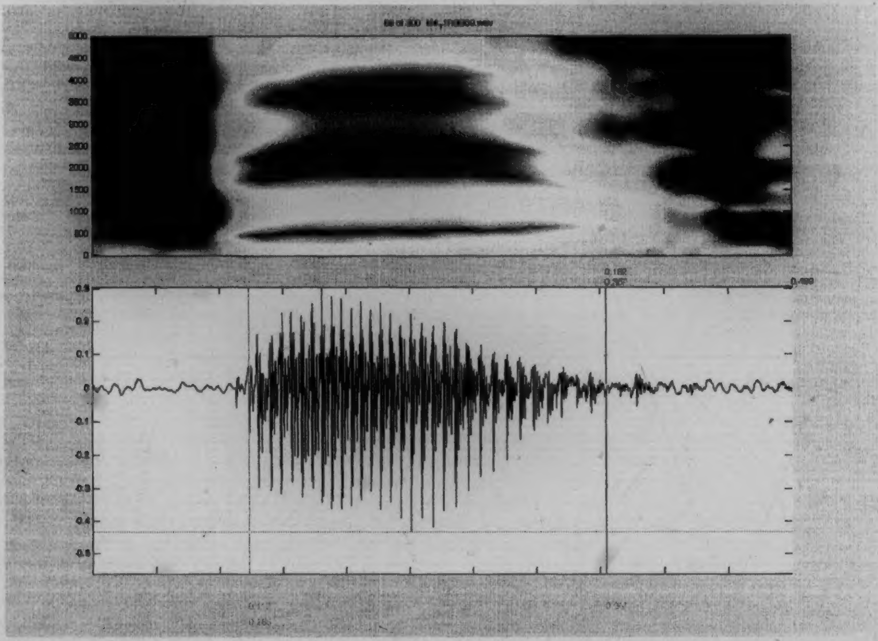


Figure 3.1. Spectrogram and waveform of English [i] with vowel boundaries marked on the waveform.

The onset of the vowel was marked as closely as possible to the first voicing pulse after the consonant release burst. The end of the vowel was not as easy to determine, particularly for the Mandarin tokens. For most of the English data, the vowel end featured a rather abrupt decrease in amplitude accompanied by the disappearance of well-defined formants. For the Mandarin productions, the amplitude and formants often tended to tail off more gradually, accompanied by evidence of glottalization or breathy voice. For these tokens, the end of the vowel was marked at the point where I could no longer audibly distinguish the vowel quality (i.e., if the breathiness was distinguishable as a specific vowel, I included it; if the breathiness contained no distinguishable vowel quality, only audible as noise, I excluded it). This seemed to be the most rational choice given my desire to include as much as could be perceived by a human listener. In some extreme cases, the latter portions of some Mandarin diphthongs were almost entirely contained within the breathy portion of the vowel. Had I marked the boundary on the basis of where the amplitude of the vowel dropped off substantially, it would have rendered the production a monophthong. Clearly, this would have had an undesirable effect on the final analysis.

After vowel boundaries were marked, an automatic formant and pitch tracker tool was used to extract values for formants and pitch for each token. These results were screened for accuracy. The Matlab program provided me with eight graphically represented alternatives based on LPC analysis at eight different frequency cut-off points, ranging from 3000 Hz to 4500 Hz. I selected a best choice from these eight alternatives through visual examination, as well as through playback of resynthesized versions of possible choices. Where necessary, I manually adjusted formant tracks where they failed to follow the correct formant. This process is illustrated in Figure 3.2 and Figure 3.3.

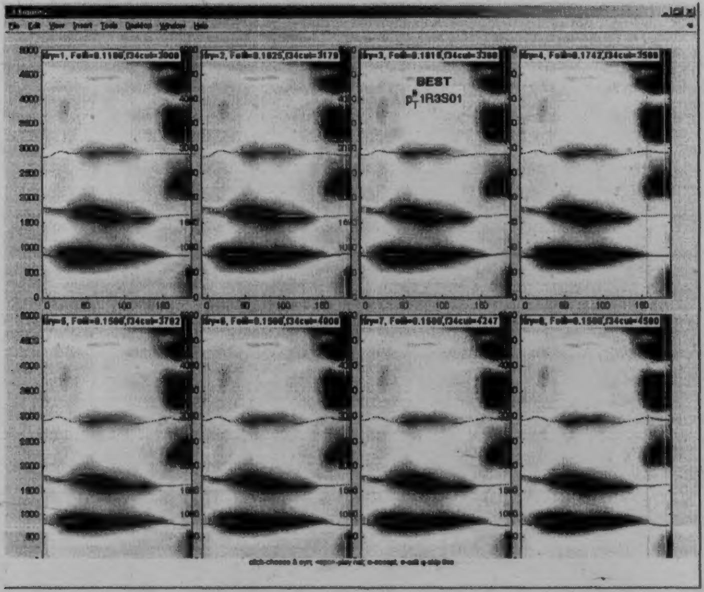


Figure 3.2. Eight alternative results of LPC automatic formant tracking based on frequency cut-offs between 3000 Hz and 4500 Hz (Successful first-pass).

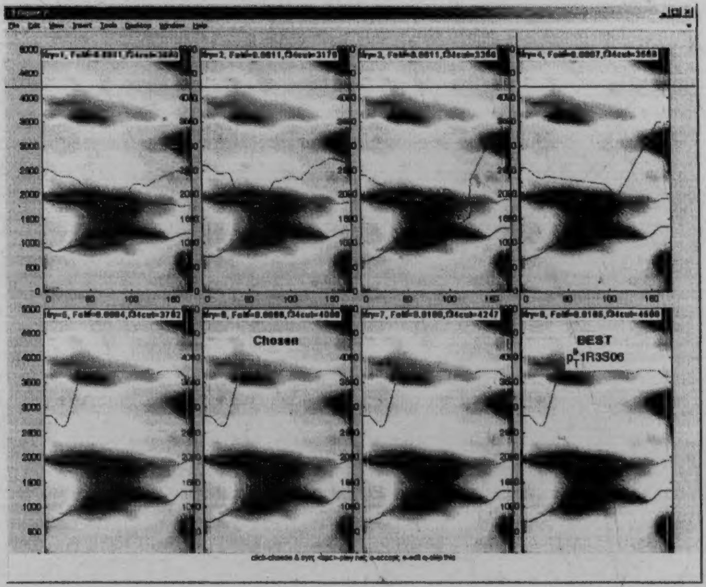


Figure 3.3. Eight alternative results of LPC automatic formant tracking based on frequency cut-offs between 3000 Hz and 4500 Hz (Unsuccessful first-pass).

Figure 3.2 illustrates analysis of a vowel recording where the automatic formant tracking algorithm was successful and multiple choices are readily acceptable. Figure 3.3 illustrates an example where the automatic formant tracking algorithm failed to accurately track all formants in any of the alternatives. In such cases, it was necessary to select the best alternative and manually adjust the formant tracking to match the visually identifiable formants. Figure 3.4 illustrates the modified formant tracks after manual correction for the best alternative from Figure 3.3.

Results of an autocorrelation pitch-tracking algorithm were also screened and corrected where possible. All formant frequency, pitch and duration values were then extracted and converted to a log scale, following Hillenbrand and Nearey (1999); a log scale provides a more accurate reflection of the human auditory system than raw measures as well as more consistent statistical properties.

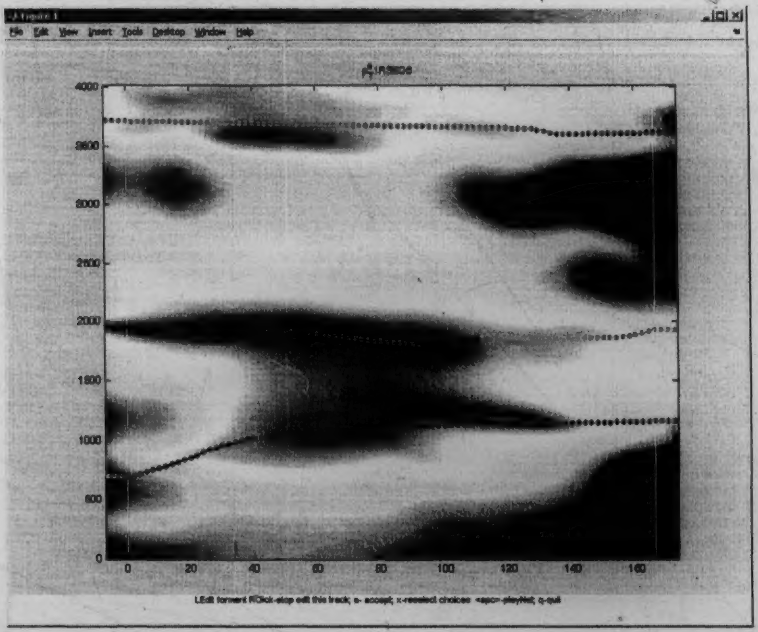


Figure 1
File Edit View Insert Tools Desktop Window Help
4000
3500
3000
2500
2000
1500
1000
500
0
0
20
40
60
80
100
120
140
160
LEdit formant RClick-stop edit this track; a- accept; x-reselect choices <spc>-play/tell; q-quit

Figure 3.4. Manually adjusted formant tracks from among alternatives previously illustrated in Figure 3.3.

For the purposes of building the pattern recognition models, I determined that I would use initial and final F1, F2, and F3 values taken from the 20% and 70% marks of each vowel's duration. This allowed me to avoid the inclusion of formant transitions from preceding consonants as well as the edges of vowel tails, which as noted were sometimes difficult to determine. The average pitch for each item was also calculated after correcting for anomalies in a limited number of F0 measurements; the pitch tracker sometimes failed to accurately track the pitch throughout the entire vowel. To adjust for this, values that suddenly dropped off to 0 Hz were omitted. The median value for all vowels produced by a given speaker was then calculated. Finally, the median value of F0 was substituted for values within each subject's vowels that were more than 3/4 of an octave above or below his/her median. This was done to insure that the mean pitch that was calculated was not overly skewed by obvious errors made by the automatic pitch tracker (e.g., a point where the pitch was calculated to be 0 Hz).

Having extracted the target values, three pattern recognition models based on discriminant function analysis were trained and tested on the English, Mandarin and Mixed (English and Mandarin) data. The Mandarin Pattern Recognition model

(Mandarin Model) contained the seven Mandarin vowel categories being examined, the English Pattern Recognition model (English Model) contained ten English vowels, and the Metamodel, treating Mandarin and English vowels as separate categories within a single system, contained a total of 17 vowel categories. Since each model was being trained and tested on the same production data, I used a round-robin cross-validation approach whereby each speaker to be tested was excluded from the training set on which his or her productions would then be tested.

In addition, in order to compare these results with those predicted by a total assimilation model where all L2 categories must assimilate to an L1 category, I also tested English productions against the Mandarin Model and Mandarin productions against the English Model.

3.2. Results

First, the relative accuracy of the L1 English and L1 Mandarin pattern recognition models were assessed to insure they accurately categorized productions in the intended categories. Results are provided in Tables 3.3 and 3.4[7].

The Mandarin Model (Table 3.3) accurately classified items 94% of the tokens when duration was excluded as a discriminating variable. Some variation in accuracy across Mandarin vowel categories is evident. However, when examining APPs by item, most misclassified tokens are recognized as having some degree of similarity to the intended category by the Mandarin Model. For example, of the eight misidentified Mandarin /ɤ/ tokens, seven had at least some probability of being the intended category (mean APP = .23; range .02 -.39). The next two most frequently misclassified categories in Mandarin showed similar patterns: all five misclassified tokens of /uə/ had some probability of being the intended category (mean APP = .21; range .01 -.42); and all four misclassified tokens of /o/ had some probability of being the intended category (mean APP = .34; range .20 -.47).

---

[7] The Mandarin Model is most accurate when duration is excluded as a variable. For the English Model, the reverse was true. Tables A1.1 and A1.2 in Appendix 1 provide full confusion matrices for Mandarin with duration as a variable and English without duration as a variable.

Table 3.3. Mandarin Model trained and tested on native speaker Mandarin productions without vowel duration as a variable.

		Vowel identified by Mandarin pattern recognition model						
		/i/	/e/	/a/	/uə/	/o/	/ɤ/	/u/
Intended Mandarin vowels repeated in response to auditory stimuli	/i/	**100**	--	--	--	--	--	--
	/e/	2.5	**97.5**	--	--	--	--	--
	/a/	--	--	**100**	--	--	--	--
	/uə/	--	--	--	**87.5**	7.5	5	--
	/o/	--	--	--	--	**90**		10
	/ɤ/	--	--	--	20	--	**80**	--
	/u/	--	--	--	--	--	--	**100**
Total Correct		94% (91% with vowel duration)						

As can be seen from Table 3.4, the English Model is also quite accurate. When duration was included as a variable, it classified tokens as the intended category 91% of the time. As with the Mandarin Model, some variation in accuracy scores across vowel categories is evident. Again, as with Mandarin, nearly all vowel tokens that were misclassified were still identified by their APP scores as having some probability of being the intended vowel. For example, eight of nine misclassified tokens of the least accurately identified English vowel, /ɛ/, had some probability of being the intended category (mean APP = .16; range .01 -.41). The next two most frequently misclassified categories in English, showed similar patterns: all eight misclassified tokens of /ʌ/ had some probability of being the intended category (mean APP, .28; range .02 - .48); and all five misclassified tokens of /æ/ had some probability of being the intended category (mean APP = .36; range .23 -.46).

Table 3.4. English Model trained and tested on native speaker English productions with vowel duration included as a variable.

		Vowel identified by English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended English vowels repeated in response to auditory stimuli	/i/	**95**	--	5	--	--	--	--	--	--	--
	/ɪ/	--	**92.5**	--	7.5	--	--	--	--	--	--
	/e/	--	--	**100**	--	--	--	--	--	--	--
	/ɛ/	--	12.5	--	**77.5**	10	--	--	--	--	--
	/æ/	--	--	--	7.5	**87.5**	2.5	2.5	--	--	--
	/ɒ/	--	--	--	--	5	**90**	5	--	--	--
	/ʌ/	--	--	--	--	5	2.5	**80**	--	12.5	--
	/o/	--	--	--	--	--	--	--	**97.5**	--	2.5
	/ʊ/	--	--	--	5	--	--	5	--	**90**	--
	/u/	--	--	--	--	--	--	--	--	--	**100**
Total correct		91% (86% without vowel duration cue)									

Having established a satisfactory level of accuracy for the Mandarin and English Models, the results of the Metamodel, excluding duration as a variable are shown in Table 3.5. It is clear that the Metamodel is far less accurate (73%) than Mandarin and English models in classifying the data according to the intended categories. Accuracy rates were slightly better (75.4%)[8] when a duration variable was included in the model. However, since the goal is to identify possible Mandarin-English interactions, and since the Mandarin Model indicated that for Mandarin speakers, duration did not serve as a useful cue, I decided to base my analysis on Metamodel results which exclude duration as a variable. Additionally, the difference in stimuli used to elicit data for each language may introduce an undesirable bias, where results reflect the relative duration of the stimuli rather than meaningful crosslinguistic differences in vowel duration.

---

[8] Table A1.3 in Appendix 1 provides a confusion matrix for the Metamodel which includes duration as a variable.

Table 3.5. Metamodel trained and tested on L1 English and L1 Mandarin productions without vowel duration included as variable. Shaded areas reflect misclassifications in opposing language.

Intended vowels produced in English or Mandarin		Vowel recognized by Metamodel																
		English										Mandarin						
		/i/$_e$	/ɪ/$_e$	/e/$_e$	/ɛ/$_e$	/æ/$_e$	/ɒ/$_e$	/ʌ/$_e$	/o/$_e$	/ʊ/$_e$	/u/$_e$	/i/$_m$	/e/$_m$	/a/$_m$	/uə/$_m$	/o/$_m$	/ɤ/$_m$	/u/$_m$
English	/i/$_e$	**65**	--	5	--	--	--	--	--	--	--	30	--	--	--	--	--	--
	/ɪ/$_e$	--	**85**	--	15	--	--	--	--	--	--	--	--	--	--	--	--	--
	/e/$_e$	--	2.5	**85**	--	--	--	--	--	--	--	--	12.5	--	--	--	--	--
	/ɛ/$_e$	--	10	--	**82.5**	7.5	--	--	--	--	--	--	--	--	--	--	--	--
	/æ/$_e$	--	--	--	7.5	**77.5**	--	10	--	--	--	--	--	5	--	--	--	--
	/ɒ/$_e$	--	--	--	--	--	**60**	10	--	--	--	--	--	30	--	--	--	--
	/ʌ/$_e$	--	--	--	--	5.1	10.3	**61.5**	--	7.7	--	--	--	15.4	--	--	--	--
	/o/$_e$	--	--	--	--	--	--	--	**60**	--	2.5	--	--	--	--	37.5	--	--
	/ʊ/$_e$	--	--	--	2.5	2.5	--	2.5	--	**67.5**	--	--	--	--	--	--	25	--
	/u/$_e$	--	2.5	--	--	--	--	--	--	--	**95**	--	--	--	--	--	--	2.5
Mandarin	/i/$_m$	27.5	--	--	--	--	--	--	--	--	--	**72.5**	--	--	--	--	--	--
	/e/$_m$	--	--	25	--	--	--	--	--	--	--	2.5	**72.5**	--	--	--	--	--
	/a/$_m$	--	--	--	--	2.5	30	5	--	--	--	--	--	**62.5**	--	--	--	--
	/uə/$_m$	--	--	--	--	--	--	--	--	--	--	--	--	--	**87.5**	5	5	2.5
	/o/$_m$	--	--	--	--	--	--	--	30	--	--	--	--	--	--	**62.5**	2.5	5
	/ɤ/$_m$	--	--	--	5	--	--	--	--	27.5	--	--	--	--	22.5	--	**45**	--
	/u/$_m$	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	**100**
Total Correct		73% without duration cue (75.4% with duration cue)																

The Metamodel confusion matrix in Table 3.5 shows clear confusion between some English and Mandarin categories, while there is very little if any confusion among others. Raw APPs also provide an interesting picture. Table 3.6 provides a percentage of English tokens for each category that had at least a .05 probability of being classified as a competing Mandarin category and the APP ranges and means for each. So, for example, in Table 3.6, of 40 English /i/ productions, 85% had at least some probability of being a member of the Mandarin /i/ category, though their APP ranged from .05 to .94. Looking at each English vowel category, we can see that many tokens have at least some probability of being classified as a similar Mandarin vowel.

Table 3.6. Percentage of English vowel tokens (n = 40 per vowel) with APPs of >.05 of being classified as a Mandarin vowel in the Metamodel.

English Vowel	Closest Mandarin Vowel	% with > .05 probability of being Mandarin	APP range (Mean)	Degree of similarity
/i/$_{e}$	/i/$_{m}$	85%	.05- .94 (.40)	high
/ɪ/$_{e}$	n/a	0%	n/a	very low
/e/$_{e}$	/e/$_{m}$	40%	.07- .99 (.36)	moderate/ high
/ɛ/$_{e}$	n/a	0%	n/a	very low
/æ/$_{e}$	/a/$_{m}$	32.5%	.08- .62 (.22)	low
/ɒ/$_{e}$	/a/$_{m}$	95%	.05- .83 (.24)	high
/ʌ/$_{e}$	/a/$_{m}$	55%	.06- .59 (.22)	low/ moderate
/o/$_{e}$	/o/$_{m}$	80%	.05- .90 (.39)	high
/ʊ/$_{e}$	/ɤ/$_{m}$	77.5%	.06- .98 (.33)	high
/u/$_{e}$	/u/$_{m}$	2.5%	(n/a)	very low

Table 3.7 provides a similar analysis for the percentage of Mandarin tokens for each category that had at least a .05 probability of being classified as a competing English category and the APP ranges and means for each.

Table 3.7. Percentage of Mandarin vowel tokens with APPs of >.05 of being classified as an English vowel in the Metamodel.

Mandarin Vowel	Closest English Vowel	% with > .05 probability of being English	APP range (Mean)	Degree of similarity
/i/$_m$	/i/$_e$	92.5%	.08 - .91 (.30)	high
/e/$_m$	/e/$_e$	47.5%	.08 - .98 (.49)	moderate/ high
/a/$_m$	/ɒ/$_e$	85%	.06 - .85 (.34)	high
/uə/$_m$	n/a	0%	n/a	very low
/o/$_m$	/o/$_e$	80%	.05 - .86 (.37)	high
/ɤ/$_m$	/ʊ/$_e$	57.5%	.08 - .92 (.46)	moderate/ high
/u/$_m$	n/a	0%	n/a	very low

Finally, Tables 3.8 and 3.9 provide confusion matrices and average APP scores for English vowels classified by the Mandarin Model and Mandarin vowels classified by the English Model[9]. These provide an indication of the types of assimilation patterns that might be expected if all items are treated as though they must assimilate to the closest category in the L1, regardless of how far they are from that category. That is, if there is no 'uncategorizable' or 'new' option available.

---

[9] Confusion matrices for the English by Mandarin model with duration and Mandarin by English without duration are provided in Appendix 1 as Tables A1.4 and A 1.5.

Table 3.8. English items tested on the Mandarin Model without vowel duration as a variable.

		Vowel identification by Mandarin Model Percentage of tokens (Average APP)						
		/i/$_m$	/e/$_m$	/a/$_m$	/uɔ/$_m$	/o/$_m$	/ɤ/$_m$	/u/$_m$
Vowel tokens produced by NSs in response to English stimuli	/i/$_e$	95 (.97)	5 (.79)	--	--	--	--	--
	/ɪ/$_e$	--	15 (.95)	--	--	--	85 (.96)	--
	/e$_e$	--	100 (1.00)	--	--	--	--	--
	/ɛ/$_e$	--	2.5 (.98)	45 (.93)	--	--	52.5 (.91)	--
	/æ/$_e$	--	--	100 (1.00)	--	--	--	--
	/ɒ/$_e$	--	--	100 (1.00)	--	--	--	--
	/ʌ/$_e$	--	--	77.5 (.96)	--	--	22.5 (.80)	--
	/o/$_e$	--	--	--	--	100 (.96)	--	--
	/ʊ/$_e$	--	--	2.5 (1.00)	--	--	97.5 (.96)	--
	/u/$_e$	--	--	--	--	17.5 (.81)	42.5 (.84)	40 (.87)

Table 3.9. Mandarin items tested on the English Model with vowel duration included as a variable.

Vowel produced in response to Mandarin stimuli	Vowel identification by English Model Percentage of tokens (average APP)									
	/i/$_{e}$	/ɪ/$_{e}$	/e/$_{e}$	/ɛ/$_{e}$	/æ/$_{e}$	/ɒ/$_{e}$	/ʌ/$_{e}$	/o/$_{e}$	/ʊ/$_{e}$	/u/$_{e}$
/i/$_{m}$	100 (1.00)	--	--	--	--	--	--	--	--	--
/e/$_{m}$	2.5 (1.00)	--	97.5 (1.00)	--	--	--	--	--	--	--
/a/$_{m}$	--	--	--	--	10 (.73)	55 (.83)	35 (.86)	--	--	--
/uə/$_{m}$	--	--	--	--	--	15 (.92)	--	7.5 (1.00)	77.5 (.96)	--
/o/$_{m}$	--	--	--	--	--	--	--	97.5 (.96)	--	2.5 (.61)
/ɤ/$_{m}$	--	--	--	5 (.95)	--	5 (.98)	--	--	90 (.98)	--
/u/$_{m}$	--	--	--	--	--	--	--	70 (.96)	7.5 (.92)	22.5 (.89)

3.3. Discussion

The results of this statistical approach to vowel measurement provide an interesting picture of monolingual Mandarin and monolingual English vowel systems, as well as their degree of similarity to each other.

3.3.1. Mandarin and English models

Although very accurate, the monolingual Mandarin and English models were not as accurate as Nearey and Assmann's (1986) model, which demonstrated a .99 correlation with human listeners. There are several possible reasons for these apparently different results. First, the sample sizes I used for my study were much bigger. In Nearey and Assmann (1986), only one production of each English vowel by four individual speakers was used, for a total of 40 tokens. In my study the sample contained two productions of each target vowel by 20 speakers of Mandarin and 20 speakers of English, for a total of 280 Mandarin tokens and 400 English tokens. The data elicitation technique used in my study may also have resulted in greater within-category variation. In my experiment, the native speaker Mandarin and English productions were elicited from linguistically naïve speakers in response to auditory stimuli. In Nearey and Assmann (1986) the productions were much more carefully elicited. The samples were a subset of tokens taken from Assmann, Nearey and Hogan (1982). In that study, productions were recorded from both an auditory prompt as well as through reference to phonetic symbols, necessitating that speakers be somewhat familiar with the phonetic alphabet. Furthermore, the authors report carefully monitoring the productions and discarding any tokens that either the experimenter or the speakers felt were inaccurate. The approach would clearly result in fewer ambiguous tokens. Another important difference between the current study and Nearey and Assmann's (1982, 1986) studies is that the targets in the previous studies were isolated vowels, whereas in the current study they were in CV frames. It was clear during my data collection that for the English speakers in particular, clear recognition of the target vowels in primarily nonce words was a difficult task.

While on the surface within-Mandarin and within-English vowel confusions may be seen as a weakness of the model, they are more likely a strength, reflecting real-world confusions. The types of confusions that are evident in these models are not unexpected. For example, the largest numbers of errors found were within English /ɛ/, where confusion with /ɪ/ and /æ/ is evident. Bohn and Flege (1992) also found that distributions of English /ɛ/ and English /ɪ/ demonstrate some overlap in the spectral dimension. Such within-language confusion patterns indicate that some categories may be more difficult to perceive in contrast with each other than are other categories where the model achieves higher accuracy scores. The fact that native English speakers produced tokens that were not unambiguous members of the intended category may indicate that these categories are inherently more difficult to learn because they are not only potentially confusable with Mandarin categories, but with other English categories. Using Eckman's (1977) terminology, they may be more marked. By marked, Eckman refers to sound categories that are less common in the world's languages. They are hypothesized to be less common because they are universally less discernable in some way than unmarked categories and therefore less likely to be incorporated into sound systems.

3.3.2. Metamodel

The results of the Metamodel analysis provide clear information regarding Mandarin and English vowel similarity. For example, we can see that Mandarin /i/ and English /i/ categories are extremely similar, although not identical. In absolute terms, Mandarin /i/ is classified as English /i/ 27.5% of the time, while English /i/ is classified as Mandarin /i/ 30% of the time. Looking at raw probabilities (as defined by APPs) of being classified as the opposing language category, English /i/ has a greater than 5% chance of being Mandarin /i/ 85% of the time while Mandarin /i/ has a greater than 5% chance of being English /i/ 92.5% of the time.

In contrast, English /u/ is extremely dissimilar to any Mandarin category. Only once is it misclassified as a Mandarin /u/. Apart from this single misclassification, none of the remaining English /u/ productions have a greater than 5% APP score on any Mandarin vowel. Other English vowels, /ɒ/, /æ/ and /ʌ/, show varying degrees of similarity to a single Mandarin category, /a/, both in terms of absolute classification and APP scores.

3.3.3. English x Mandarin and Mandarin x English models

The results of the English x Mandarin and Mandarin x English vowel classification models provide a somewhat different picture of what would happen if all L2 categories were forced to assimilate to an L1 category. These models would predict confusion in the L2 whenever more than one L2 category assimilates to a single L1 category. For example, tested against the Mandarin Model, 85% of English /ɪ/ productions assimilate to Mandarin /ɤ/. Since English /ɛ/, /ʊ/ and /u/ also assimilate to the same Mandarin category, this seems to predict possible confusions between English /ɪ/ and these other English categories. Such a prediction seems unlikely to be correct, given that the absolute phonetic distance from English /ʊ/ and /u/ to English /ɪ/ and /ɛ/ is quite large. Such an unlikely prediction demonstrates the risk associated with assuming all L2 production tokens *must* assimilate to an L1 category. In reality, these English x Mandarin and Mandarin x English models only provide general information regarding what opposing language category each production token is closest to in absolute terms; they do not indicate the precise degree of similarity. Hence, while a production of English /ɪ/ may be categorized by the Mandarin Model as Mandarin /ɤ/, that does not entail that the production is as close to Mandarin /ɤ/ as are productions of English /ʊ/ and /u/.

3.3.4. Predictions for Mandarin L1 learners of English

The results of the absolute classifications and raw APPs from the Metamodel in this experiment allow us to make a set of very explicit predictions concerning perception and production of English vowels by Mandarin L1 speakers. Although this approach may be used to make SLM based predictions about ultimate discrimination between Mandarin and English vowels (the sort of task conducted for Japanese and English consonants in Guion et al. [2004] where learners were asked to discriminate between similar and dissimilar Japanese-English contrasts), this dissertation only tests predictions in terms of identification and production of English vowels. Since absolute identification does not require nativelike perception of 'similar' categories but only the ability to discern some difference, for the perceptual identification portion, PAM provides an appropriate conceptual starting point for making predictions. For the L2 production data where gradient measures of acoustic similarity to the intended target is possible, the SLM is a more appropriate gateway.

Predictions regarding L2 English vowel identification in perception

When a Mandarin category is very similar to a single English category, identification of that English category should be strong because something resembling PAM's direct category assimilation occurs. To operationalize similarity in Metamodel terms, I define *statistically similar* vowels as those that have a greater than .05 APP of being a member of the competing vowel category (refer back to Tables 3.6 and 3.7). When more than one English category is statistically similar to a single Mandarin category, single category assimilation will result. When an L2 English category is not statistically similar to any Mandarin category, something PAM might call an uncategorizable speech sound, it may or may not be difficult for beginning learners to perceive and identify that English vowel. For the purposes of this discussion, such uncategorizable L2 English vowels are in Metamodel terms recognized as *statistically new* vowels. I define statistically new vowels as those vowels that have a less than .05 APP of being a member of a competing L2 category. In fact, most sounds that have any probability of being a member of a competing language category have a much greater than .05 APP of being so. In contrast, almost all statistically new sounds have a less than .001 APP of being a member of a competing language category. If a statistically new English category is unambiguous with regards to other sounds within the English Model, we should expect L2 English learners to develop an ability to identify that category relatively quickly. In contrast, if the statistically new English category is confusable with other English vowels within the English Model, we should expect Mandarin learners of English to have some difficulty perceiving it vis-à-vis the other competing English vowel(s).

Following from the Metamodel, we can thus summarize predictions regarding English vowel learning by Mandarin speakers as follows: English /i/, /e/, /o/, /ʊ/ are most statistically similar to Mandarin /i/, /e/, /o/ and /ɤ/, respectively, and will therefore be primarily assimilated to these Mandarin categories. Consequently, we should expect strongest identification rates for those English categories. In addition /ɒ/, /æ/ and /ʌ/ are all statistically similar to a single Mandarin /a/ category. Therefore, we should expect confusion among these categories, although /ɒ/ will be identified more accurately than the other two statistically similar English categories because it is more statistically similar to Mandarin /a/ than are /æ/ and /ʌ/. According to PAM, learners' ability to discern some differences in the goodness of fit of each of the three English categories to a single Mandarin category will result in slightly different responses to each English category. The English vowels /ɪ/, /ɛ/ and /u/ are statistically new categories for Mandarin learners of English. Among these three, /u/ will be relatively easy to discriminate because it has good within-English distinctiveness. In contrast, /ɪ/ and /ɛ/ may be more difficult to discriminate because of within English ambiguity.

Predictions regarding L2 English production

While crosslinguistic statistical similarity between Mandarin and English will aid in identification of some English vowel contrasts, statistical similarity can still have a negative effect on production. Following claims of the SLM, the Metamodel predicts that the English vowels /i/, /e/, /o/, /ʊ/ and /ɒ/ will often be recognized in production as members of the statistically similar Mandarin category. The slightly less statistically similar English vowels /æ/ and /ʌ/ will also sometimes be recognized in production as members of the Mandarin category /a/, although they have a better chance of developing as independent categories. In contrast, /ɪ/, /ɛ/ and /u/, because they are statistically new vowels for Mandarin speakers of English, are predicted to be least Mandarin-like in production. As it stands, the Metamodel makes no predictions regarding rate of acquisition of these statistically new vowels. Since they are not bootstrapping on Mandarin categories, we should assume that their correct articulation has to be learned and therefore initial inaccuracy may be the norm. In acquiring these statistically new categories, learners must first be able to notice that some NS productions of these categories are sufficiently different from L1 Mandarin categories to warrant establishing a new category. Subsequently, the exact parameters of that category must be determined over time through some sort of statistical learning (e.g., clustering) of productions of that category in the learners' input. As I suggested in earlier work (Thomson, 2003), it is possible that some less than prototypical input may initially serve as the basis for these statistically new categories, especially if the learners are initially only able to notice differences between non-prototypical productions of the statistically new category and existing L1 categories. Ultimately, when the statistically new category begins to emerge in the learners' L2 production, errors are unlikely to be classified as Mandarin vowel categories since, however inaccurate, the learner is aware that these are not similar to Mandarin vowels.

In a later chapter, specific predictions resulting from the Metamodel analysis of vowel similarity reported in this chapter will be applied to L2 vowel identification and production data from a training study outlined in the next chapter. This later analysis will help to further explicate differences between traditional SLM and PAM claims and those quantified by the Metamodel approach. For the sake of succinctness and to maintain a clear connection with the Metamodel's origins in claims made by the SLM, I will continue to use the terms, 'similar' and 'new' to refer to relationships between L1 and L2 sounds. However, as mentioned earlier, in the context of the Metamodel, these terms should always be understood to mean *statistically* similar and *statistically* new.

Chapter 4. Training and its effect on perception and production

The training experiment reported in this section was motivated by findings from previous research, which have provided some understanding of factors that contribute to the development of L2 perceptual categories. In particular, individual differences in L2 phonological attainment are consistently correlated with the degree of experience learners have with the L2 (Flege, Bohn, & Jang, 1997; Flege, Frieda, & Nozawa, 1997; Yamada, 1995). Pedagogical and experimental interventions aimed at encouraging more rapid establishment of L2 categories have often resulted in differential success (Derwing, Munro & Wiebe, 1997, 1998; Jamieson & Morosan, 1986, 1989; Logan, Lively & Pisoni, 1993; McCandliss, Fiez, Protopas, Conway, & McClelland, 2002; McClelland, Fiez, & McCandliss, 2002). When such interventions have contributed to improvement, it seems reasonable to conclude that this is the result of learners having been provided with more effective experience with L2 phonetic input than is normally afforded them in natural contexts. While improved experience may simply be the result of greater frequency and intensity of exposure, it may also be related to the quality of the input itself. By this, I mean the degree to which the input is salient or noticeable to the learner. Providing training that enhances the learner's ability to attend to L2 phonetic input in a way that maximizes its potential for incorporation into the developing interlanguage system is critical.

In the training study presented here, I am interested in testing the global effects of different types of instruction as well as testing the predictions of PAM and SLM based on the results of the Metamodel's determination of Mandarin-English crosslinguistic similarity. The study and results will be presented in this chapter. Discussion in terms of concepts related to PAM and SLM will be dealt with in Chapter 5.

The purpose of this study is to determine what effect three types of perceptual training will have on Mandarin speakers' identification and production of English vowel categories. One group, henceforth called the Long Vowel Training (LVT) group was trained on CV stimuli in which the vowel portion had been artificially lengthened to provide a longer duration during which learners might be able to detect important acoustic information. A second group, henceforth called the Select Vowel Training

(SVT) group was trained on a subset of naturally produced English CV tokens that were determined to be less Mandarin-like than other CV tokens from the same set of speakers. A third group, henceforth called the Deselected Vowel Training (DVT) group was trained on the opposing subset of naturally produced English CV tokens, those that were determined to be most Mandarin-like. This third condition was intended to provide some control for the first two conditions. If either of the first two types of training were found to have an effect specific to the training type, we would expect the effect of the third condition to be smaller, since these vowel training stimuli were neither artificially lengthened, nor as distinct from Mandarin categories.

A traditionally defined control group was not used because of the nature of the testing procedure; first, it was not possible to test a control group's ability to identify English vowels without them being familiar with the training paradigm. In addition, participants were asked to volunteer and were therefore limited in number. Furthermore, all wanted to benefit from training. While not ideal, I do not believe the lack of a traditionally defined control group to be a major shortcoming, since the effect of training can be contrasted with other studies, including Derwing, Munro & Thomson (2003) that demonstrate that the naturalistic development of speech perception and production require a long period of time and substantial exposure. Therefore, if more immediate changes are detected over the relatively short duration of this study, we can safely assume that they are the result of training. In addition, a delayed post-test will provide an additional form of control. If, in the absence of training, learners continue to improve over a period of time that is similar in length to the training period, we would have to conclude that training did not *necessarily* cause the initial improvement. On the other hand, there is a possibility that learners might continue to improve if the initial training provides them with a foundation for L2 English categories that could then be strengthened through naturalistic input. However, continuing improvement as the result of naturalistic input seems unlikely over such a short amount of time. Hence, we should expect that if training has an effect, learners should not also demonstrate measurable improvement in the month after training has been concluded.

4.1. Research Questions

This training experiment will address the following five research questions:

Perception

1. When learners are presented with different types of training in phonetic contrasts, what global (general) effects are there on the development of their L2 English vowel identification ability?
2. Does identification training in one CV context transfer to identification ability in another CV context?
3. To what extent does the effect of identification training transfer to new tokens produced by a familiar voice and new tokens produced by a new voice?
4. Will evidence of improvement in perception still be detectable one month after training is completed?

Production

5. To what degree does identification training result in improvements in production?

4.2. Method

4.2.1. Participants

Twenty-six adult (*M* age = 36.12 years, range = 27 – 50 years) Mandarin L1 speakers (17 women, 9 men) were recruited from a local ESL program, where they continued to receive general ESL instruction for the duration of the study. Before beginning their ESL classes, the English proficiency of all participants had been assessed as being between Benchmarks 1 and 3 in listening and speaking skills, according to the Canadian Language Benchmarks Assessment tool (they were defined as beginners). Because they had previously studied English in China, some were able to read and write in English at levels beyond their listening and speaking skills. The participants had been studying beginner level ESL in Canada for an average of 4.1 months (range = 1- 13 months) at the time they volunteered for this research. Their ESL classes included little or no explicit pronunciation instruction. All participants were well educated immigrants to Canada, and most had arrived within the previous year (*M* LOR = 11.6 months, range

= 4 - 48 months). Four who had arrived more than a year earlier reported having had little interaction with Canadians outside the local Mandarin community and no self-reported exposure to English on a daily basis prior to enrolling in the ESL program. Finally, all the participants reported having normal hearing.

The first twenty-two participants were randomly assigned to one of the first two training conditions (LVT or SVT), while the final four were assigned to the third condition (DVT). This smaller sample size in the DVT group resulted from an inability to recruit more participants. I determined that keeping the first two groups to an adequate size was more important than having an equal number in the DVT group[10]. All participants were asked to attend a number of testing and training sessions over the course of 3 – 4 weeks, during their lunch break or after their scheduled classes were finished for the day. On average, participants completed the training and testing component of the study in 21 days (range = 17-27). In order to facilitate one-on-one production testing, and introduction of the training program, the days on which learners began training was staggered, requiring a total of six weeks on-site at the ESL program.

Of the 26 initial participants, 18 (*M* age = 37.15 years, range = 30 – 50 years; 12 women, 6 men) were able to participate in a delayed post-test approximately one month later (*M* = 30.9 days; range = 26 –34 days). This subset of participants had been studying ESL for an average of 4.17 months (range = 1- 13 months) at the time they began this study and had been in Canada for an average of 11.5 months (range, 4 - 48 months). As mentioned previously, those who had arrived earlier reported having little if any interaction with Canadians outside their Mandarin community and no exposure to English on a daily basis prior to enrolling in the ESL program. Participants were paid a small honorarium for an initial production pretest as well as for their production post-test and a delayed identification post-test, but not for the training phase. Table A2.1 and A2.2 in Appendix 2 provide details for each participant by training condition, for the entire group and the delayed post-test group respectively.

---

[10] In all, 30 participants initially volunteered, 12 for each of the first two conditions and six for the third condition. One participant's data from the LVT group was excluded because, after training began, I discovered that she was Taiwanese and therefore did fit the study's requirements. I did, however, allow her to finish training, despite excluding her data. One participant from the SVT group withdrew after the first session because she found employment. Two participants from the DVT group withdrew, one after the first session and one after the fourth, both citing time constraints as their motivation. Initially, I had planned for the DVT group training condition to have the same number of participants as the other two groups however, given constraints on the number of participants who volunteered, only four participants finished the third training condition. Furthermore, upon preliminary analysis of the incoming results from the first two groups and the initial four in the DVT group, it was determined that even with 11 participants in the DVT group, it was very unlikely that I would be able to detect any effect compared to the other groups. Since this was the case, I felt, for practical and ethical reasons it would not be appropriate to continue seeking volunteers at a future date and/or location.

4.2.2. Stimuli

Training stimuli

The training stimuli were derived from the original 400 native English speaker /bV/ and /pV/ productions collected for use in the crosslinguistic similarity study reported in the previous chapter. These syllables are ideal for this training experiment because many are quite similar to real Mandarin words, making them good candidates for assimilation to L1 categories. In contrast, only eight are real English words, and with a few exceptions, those that are real English words are relatively low frequency. This decreases the possibility that previously established misperceptions of English at the lexical level will interfere with learning.

For the LVT group, all 400 recordings of the native speaker English /bV/ and /pV/ productions were modified using Praat[11], such that the vowel portions (defined as the onset of voicing after the consonant release burst until the last voicing pulse) in each of the 400 English CV productions were doubled in length. Praat employs a Pitch Synchronous Overlap Add (PSOLA) method which, simply defined, has the effect of repeating each voicing pulse in the original sound file, resulting in a relatively natural sounding but lengthened version of the original production. The resulting waveforms were saved as new sound files. The lengthened vowel stimuli were used in an attempt to provide a greater duration during which learners could discern the phonetic information that is important for vowel identification.

---

[11] Praat is a freeware phonetic analysis and manipulation program downloadable from www.praat.org

The second set of training stimuli, for the SVT group, was a subset of the 400 naturally produced /bV/ and /pV/ recordings. Using the same measurements from which the statistical pattern recognition models were derived (i.e., F0, F1, F2 and F3), Mahalobinis Distance (MD) scores from Mandarin categories were calculated for each of the English production tokens. As mentioned in Chapter 2, these scores provide a measure for assessing the absolute distance a given production token is from any category's centre. Each category's centre is calculated on the basis of all production tokens of that category. We should expect that L1 productions should have relatively low MD scores from their intended category. However, the same scores can also be used to define how close a production in one language is to a category in another language. So, for example, if an English /i/ is found to have an MD of 3.456 from Mandarin /i/, we can conclude that, in absolute terms, it is more Mandarin-like than an English production of /i/ that has an MD score of 18.543 from Mandarin /i/. Within each subset of 40 tokens for each English vowel category, those 20 that had the highest MD scores from any competing Mandarin category (i.e., they were furthest from Mandarin categories) were chosen as training stimuli, resulting in a total of 200 stimuli (20 for each vowel). For the DVT group, the remaining 200 natural stimuli tokens that had the lowest Mahalanobis distance scores (i.e., those that were more Mandarin-like) were used.

Production testing stimuli

The production test stimuli comprised 80 CV targets. The first 60 were spoken by a female native speaker (Voice 1) whose voice was not used in the training stimuli and included all 10 target English vowels, /i, ɪ, e, ɛ, æ, ɒ, ʌ, ʊ, o, u/ produced in /bV, pV, zV, sV, gV, kV/ syllables in the order listed[12]. The last 20 items were novel /bV, pV/ productions spoken by a male speaker (Voice 2) whose voice was also used in the training stimuli. All target items were recorded in the carrier phrase "The next word is ____." These 80 items were recorded onto CD for presentation purposes.

The vowel portion from each of the 80 CV syllables used for testing stimuli was measured in the same manner as were the L1 Mandarin and L1 English vowel data described in the previous chapter. Measurements taken included F1, F2 and F3 values from the 20% and 70% portions of each vowel token's length, F0 and duration. These measurements were used to test each stimulus CV production against the English and Metamodels trained in the previous chapter. Recall that the English Model was trained on English vowel production data from 20 English L1 speakers, while the Metamodel was trained on English vowel production data from the same 20 English L1 speakers in addition to Mandarin vowel production data from 20 L1 Mandarin speakers[13]. Results of testing each test stimulus CV against the English Model with vowel duration included as a variable are provided in Table A3.1 - 3.4, in Appendix 3. In total, 99% of the stimuli were correctly identified. Results indicate that all of the /b, pV/ syllables produced by stimulus voices were recognized as the intended vowel category; all /z, sV/ stimuli which were only produced by Voice 1 were recognized as the intended vowel; finally, all but one /g, kV/ stimuli which were also only produced by Voice 1 were recognized as the intended vowel. The single error was in response to Voice 1's production of /gʌ/, which was recognized as /gɛ/. When vowel duration was excluded as a variable from the English Model (refer to Appendix 3, Tables A3.5 –A3.8), 94% of the stimuli were correctly identified. Among the errors, Voice 2's production of /bæ/ was recognized as /bɛ/; his production of /pɒ/ was recognized as /pʌ/. Stimulus Voice 1's production of /gɛ/ was recognized as /gɪ/; /gʌ/ was recognized as /gɛ/; her production of /zɒ/ was recognized as /zʌ/. These confusions indicate that in the spectral dimensions, although relatively rare, some ambiguity is present due to some overlap in spectral properties. Finally, each stimulus CV was tested against the Metamodel, excluding vowel duration as a variable (refer to Appendix 3, Tables A3.9 – A3.12). These results indicate that the degree of similarity to Mandarin categories for each stimulus CV varies somewhat across context and stimulus voice.

---

[12] The first 20 items were the same as those used to elicit NS English productions for the English and Metamodels in the previous chapter.

Perception testing stimuli

Three sets of testing stimuli were used. The first set (Generalization Test) comprised items not used in training. These were the same 80 CV tokens used in the production test just described. For the identification test, however, they were extracted from their sentence frames. A second set (Lengthened Vowel Test) comprised all 400 vowel-lengthened stimuli. The third set (Natural Vowel Test) comprised the entire set of natural stimuli (including those with both higher and lower Mahalanobis distance scores.

---

[13] Since Voice 2 was used to provide input to the original English Model, from which training stimuli were then derived, this speaker's original productions were excluded from the English Model trained to evaluate CV test stimuli produced by Voice 1 and Voice 2. The remaining 19 L1 English speaker voices were the same as those used in the original English Model described in Chapter 3.

4.2.3. Procedure

The order of all training and testing procedures is summarized in Table 4.1. Details about each phase follow the summary presented in the table. The implementation of all phases took approximately 3-4 weeks. I could not control when volunteers were unable to attend and therefore could not maintain a stricter time schedule. In addition, participants did not receive training or testing on weekends. This meant that some training and testing inevitably straddled weekends. Although I tried to avoid this, often it was beyond my control. My only stipulation was that each participant attempt to attend training or testing sessions on at least four separate days each week. With a few exceptions, this was accomplished.

Pre and post-training production recordings

Both before and after receiving training, the participants' English productions were recorded in a quiet room. The 80 CV elicitation stimuli were presented via headphones. As described earlier, each item was presented in the carrier phrase, "*The next word is* _____". Participants were asked to respond by repeating the word in the carrier, "*Now I say* _____". This procedure was repeated twice for each participant in order to obtain two recordings of each item, for a total of 160 items. After recording each of the speakers' productions, the target syllables were extracted from the resulting wave files, down sampled to 22.055 kHz, normalized across tokens at peak amplitude, and then saved as separate sound files for each syllable. This was only done for the LVT and SVT groups because the small sample size of the DVT group suggested there was insufficient power to expect a significant result.

Table 4.1. Order of training and testing phases with approximate timeline in parentheses.

Phase	*Brief description*
Production Test 1	Pre-training recording of participants' productions from elicited imitation of English CV syllables. 80 items: 60 Voice 1; 20 Voice 2 (Day 1)
Training Demo	Initial demonstration of training program (Day 1 or Day 2)
Training Phase 1	Four training sessions, 200 items per session (Completed over approximately 4 – 7 days)
Lengthened Vowel/Natural Vowel Test 1	LVT on Lengthened Vowel Test, SVT and DVT on Natural Vowel Test (1-3 days after completing Training Phase 1)
Generalization Test 1	Both groups tested on identification of non-training items in multiple CV contexts 80 items: 60 Voice 1; 20 Voice 2 (1-3 days after completing previous test)
Training Phase 2	Four training sessions, 200 items per session (Completed over approximately 4 – 7 days)
Lengthened Vowel/Natural Vowel Test 2	LVT on Lengthened Vowel Test, SVT and CVT on Natural Vowel Test - 400 items (1-3 days after completing Training Phase 2)
Lengthened Vowel/Natural Vowel Test (Alternate)	LVT on Natural Vowel Test, SVT and CVT on Lengthened Vowel Test - 400 items (1-3 days after completing previous test)
Generalization Test 2	Both groups tested on discrimination of new items in multiple CV contexts 80 Items: 60 Voice 1; 20 Voice 2 (1-3 days after completing previous test)
Production Test 2	Post-training recording of participants productions from elicited imitation of English CV syllables 80 Items: 60 Voice 1; 20 Voice 2 (Within a day of completing generalization test)
Delayed Post-test	All participants on Natural Vowel Test and Generalization Test (Approximately one month after previous tests)

Demo session

On either the day of the initial production recording or the following day, participants were provided with a demonstration of how the training program worked. Training was implemented with a computer program written by Terrance Nearey using Matlab. The design of the identification training program was influenced by one described in Guion and Pederson (2007b) that was used to train English learners of Hindi sound contrasts. In Demo mode, learners were shown a series of 10 images (see Figure 4.1) of international nautical flags and were told that each flag would be associated with a particular English vowel spoken in a CV context. Nautical flags were used as labels for categories because I could be relatively certain that learners had no prior experience with these highly distinctive symbols. In contrast, using orthographic representations might have interfered with learning, particularly if faulty perceptions were already associated with particular orthographic or phonetic symbols. Despite being new to the learners, nautical flags can be easily differentiated on the basis of patterns and colours, and therefore learning to recognize them is not particularly difficult. In addition, the relative location of each flag on the screen did not change. The use of arbitrarily assigned non-orthographic characters was employed by Guion and Pederson (2007b), although they do not describe the explicit nature of the characters they used.

Using a single female speaker's /bV/ productions, vowel categories and their corresponding flags were introduced incrementally in pairs, before being contrasted with previously introduced items. Each item was first presented aurally. After a one-second delay, the corresponding flag flashed twice on the screen, followed by another aural repetition of the item. The learner was then asked to click on the flag representing the item just displayed. After clicking on the item, followed by another one-second delay, the next item was presented. Items within each pair were randomly played three times before moving on to the next pair.

After the first two pairs were heard, all four items within those pairs were again presented randomly, three times each, before the remaining pairs were introduced in a similar fashion. Since items were introduced multiple times within pairs and subsequently presented randomly within a larger set of items, learners were told that when they recognized an item, they should attempt to move the mouse pointer to the item they heard before it began flashing. This allowed them to begin establishing connections between the sound category and the corresponding symbol. The demonstration mode was designed to present exactly the same number of repetitions (n=12) for each item. The exact order of category introduction and repetition is provided in Table A4.1 in Appendix 4. The demo session, as with all subsequent training and vowel identification test sessions, was conducted in a computer lab at the ESL program site. Auditory stimuli were presented via headphones.

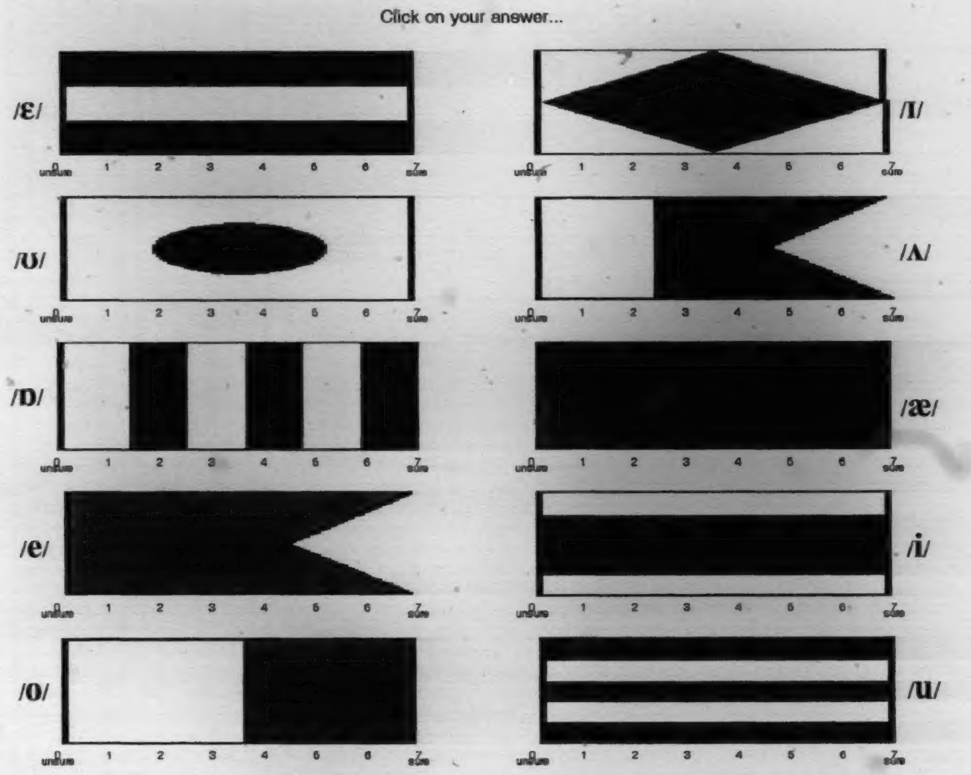

Click on your answer...

Figure 4.1. Screenshot of training program with ten images, each corresponding to the English vowel indicated to the left or right side of the image. Vowel labels are for information only and were not provided in the training program.

Training sessions

While the structure of training sessions was the same for all groups, the nature of the training stimuli differed. During training, learners were asked to continue learning to associate each image with a particular English vowel category that would be presented aurally in either a /bV/ or /pV/ frame by a variety of native speakers. During each training session, learners heard 200 randomly presented syllable tokens (20 of each English vowel) taken from the set of 400 English syllable tokens (40 of each vowel) previously described. The LVT group training sessions alternated between lengthened versions of the 200 /bV/ productions and lengthened versions of the 200 /pV/ productions. For the SVT group, every session comprised the set of 200, natural /bV/ and /pV/ tokens that had been selected on the basis of their relative distance (based on MD scores) from Mandarin categories. Finally, for every session, the DVT group heard the 200-token subset of natural vowel tokens that were closest (based on MD scores) to Mandarin categories. After hearing each production, learners were asked to click on the symbol that represented the vowel they heard and were given feedback on the accuracy of their response. If they made the correct choice, the flag flashed twice and they heard a confirmation beep, and then, after a 500 ms interval, heard the next item. If they made an incorrect choice, they heard a negative beep followed by a double flashing of the correct image, while the same aural stimulus was played a second time. They were then required to click on the correct image and continued to the next item after a 500 ms interval. The learners' first responses to each training item were automatically saved to a text file.

Vowel identification testing sessions

After four training sessions and again at the end of training, four training sessions later, learners were asked to take a number of performance tests. I chose to conduct the first identification test after four training sessions because I could be relatively certain that participants had learned the task and errors would therefore be related to perceptual difficulty, rather than lack of task familiarity. In the early training sessions, learners were still making apparently random errors on vowels that I assumed they would have no difficulty identifying because of their similarity to Mandarin (e.g., /i/). This indicated that they were still learning the associations between sound categories and symbols.

The method used for testing was nearly identical to that used for training with one important distinction - learners were not provided with feedback as to the accuracy of their responses. Instead, they heard an auditory stimulus and were required to make a choice from the ten nautical flag symbols. As in training, 500 ms after making their selection, the next item was presented. The learners' responses and ratings for each test item were automatically saved to a text file.

The test sets differed by training group as follows:

Generalization Test

At the midpoint and endpoint of training, all participants were asked to complete the identification test of 80 CV items on which they had not been trained. As described earlier, these test stimuli were CV productions extracted from the production elicitation stimuli as described earlier, including target vowels produced in CV contexts.

Lengthened Vowel/Natural Vowel Tests

For the LVT group, the first test after four training sessions (mid point) comprised the entire set of 400 lengthened vowel productions, both /bV/ and /pV/ syllables. For the SVT and DVT groups, the mid training test set comprised the entire set of 400 naturally produced tokens, both those 200 that were selected as being less Mandarin-like as well as those 200 that were deemed more Mandarin-like. After training was complete all groups took both the lengthened vowel and natural vowel tests.

Delayed post-test

For the delayed post-test, the 18 remaining participants (nine from each of the LVT and SVT groups) were tested on the Generalization Test as well as the Natural Vowel Test.

4.2.4. Data analysis

Vowel Identification Data

All identification response data for training sessions and tests were transferred into SPSS and organized by relevant variables such as group, time, test, and stimuli characteristics. A number of statistical tests were conducted to provide answers to the research questions posed at the beginning of this experiment. Finally, learner confusion patterns between English categories were used to evaluate the accuracy of predictions made by my Metamodel approach to PAM.

Vowel Production Data

All 7040 recordings of the L2 speakers' syllable productions from the LVT and SVT groups at Time 1 and Time 2 were analyzed using the same approach described for analysis of English and Mandarin vowels in the previous chapter. In brief review, using acoustic analysis software designed with Matlab, first vowel boundaries were marked. Next, automatic LPC formant tracking was used for a first pass at tracking F1, F2 and F3. These results were then manually screened and manipulated where necessary. As before with the Mandarin L1 tokens, a very large proportion of items needed to be manually adjusted because many vowel endings trailed off into breathiness or irregular voicing, a feature that was not as frequently found in the L1 English data. As with the previous vowel recordings, I marked the end of the vowel at the point where I could no longer audibly distinguish the vowel quality (i.e., if the breathiness was distinguishable as a specific vowel, I included it; if the breathiness contained no distinguishable vowel quality, and was only audible as noise, I excluded it). After making manual adjustments, the final results of the automatic formant and pitch tracking procedures were saved for analysis. As with the previous production data, all formant frequency, pitch and duration values were then extracted and converted to a log scale. The average pitch for each item was calculated after correcting for errors in some F0 measurements by replacing with that subject's median value (across all vowels) any values that fell more than 3/4 of an octave above or below the median. In the end, only five of the 7040 tokens failed to be measured using this procedure. These missing values were replaced with values from the same participant's second repetition of the same item. Finally, duration, mean F0 and values for F1, F2 and F3 taken from the 20% and 70% points of each vowel token were used as 'new' cases to be tested against the English Model and Metamodel using discriminant analysis. For the English Model, these results reflect the number of L2 production tokens that were classified as the intended English vowel category or other English vowel categories. Further statistical tests were conducted on these results to determine if differences existed in terms of training group, time or stimuli characteristics.

4.3. Results

4.3.1. Identification tests

Generalization test

The results of the vowel identification test on non-training items are presented first since they provide a consistent measure of performance across both groups and time.

Mean correct identification rates for pooled vowels by CV context, speaker and time are provided in Table 4.2. Identification rates by vowel category are provided in later figures. Differences in performance comparing /b, pV/-/g, kV/-/z, sV/ contexts for Voice 1 are treated separately from comparisons of differences in performance in response to each voice in the /b,pV/ context; Voice 2 stimuli only included the /b,pV/ context.

Table 4.2. Mean % correct vowel identification scores and standard deviations on Generalization Test by CV context, Stimulus Voice, Training Group and Time.

		/b, pV/ Voice 1			/g, kV/ Voice 1			/z, sV/ Voice 1	
Group	LVT (n=11)	SVT (n=11)	DVT (n=4)	LVT (n=11)	SVT (n=11)	DVT (n=4)	LVT (n=11)	SVT (n=11)	DVT (n=4)
Time 1 %	69.55	67.73	85.00	61.82	61.82	71.25	56.36	58.64	66.25
SD	*19.68*	*18.35*	*7.07*	*18.61*	*20.03*	*17.76*	*16.45*	*17.62*	*11.09*
Time 2 %	79.09	83.64	80.00	72.73	65.91	71.25	64.09	60.00	73.75
SD	*17.72*	*9.77*	*4.08*	*15.55*	*13.75*	*2.50*	*18.95*	*15.65*	*11.08*

		/b, pV/ Voice 2		
Group	LVT (n=11)	SVT (n=11)	DVT (n=4)	Average across groups and CVs
Time 1 %	74.55	70.45	77.50	68.41
SD	*15.73*	*16.50*	*11.90*	*15.90*
Time 2	78.64	80.00	85.00	74.51
SD	*13.43*	*10.25*	*00.00*	*11.06*

Responses to Voice 1 stimuli in /b, pV/, /g, kV/ and /z, sV/ contexts

A three-way partially repeated measures ANOVA was computed with Time (2 levels) Consonant (Voice 1 only, 3 levels) and Vowel (10 levels) as repeated measures, and Training Group (LVT, SVT and DVT) as a between-subject factor. Because Mauchly's Test of Sphericity was significant for Vowel, corrected Huynh-Feldt measures are reported. Comparison of effects found using Huynh-Feldt versus other corrected, uncorrected and multivariate Wilks' Lambda measures indicate similar results (See Table A5.1 in Appendix 5). Significant differences were found for Time [$F(1, 23) = 5.163$, $p <$

.05], Consonant [*F*(4, 46) = 26.322, *p* < .01] and Vowel [*F*(7.545, 173.528) = 18.665, *p* < .01]. A significant Consonant x Vowel interaction was also detected [*F*(16.110, 370.539) = 4.502, *p* < .01], however, the effect size for this interaction was very small (partial Eta squared = .164). Differences in mean identification scores over time for each Vowel x Consonant combination in response to the Voice 1 stimuli are illustrated in Figures 4.2 to 4.4; mean results are pooled across training groups.

No other significant interactions between within-subject factors or with Training Group were found. Post-hoc HSD Tukey tests on the effect of Consonant showed a significant difference in correct identification rates at Time 1 for Voice 1 /b, pV/-/z, sV/, but not other contrasts; at Time 2, significant differences in identification rates were found for Voice 1 /b, pV/-/z, sV/ as well as /b, pV/-/g, kV/ contrasts. All significant differences in the effect of Consonant were in favor of the /b, pV/ contexts. Examining the relative degrees of improvement between /b, pV/ compared to /g, kV/ and /z, sV/ for each vowel (see Figures 4.2 – 4.4) , it appears that although in the /b, pV/ context the identification rates of all vowels are showing a trend toward improvement, in the other two contexts, two vowels, /ɪ/ and /ɒ/, show no improvement or actually get worse over time.

The Pearson correlation between Time 1 and 2 for mean vowel identification scores in response to the Voice 1 /b, pV/ stimuli is *r* = .98; /g, kV/ stimuli, *r* = .97; and /z, sV/ stimuli, *r* = .95, indicating that the relative identification rates across vowels remained the same. In other words, those vowels that had relatively weaker identification scores at Time 1, though perhaps improving, were still relatively weak at Time 2.

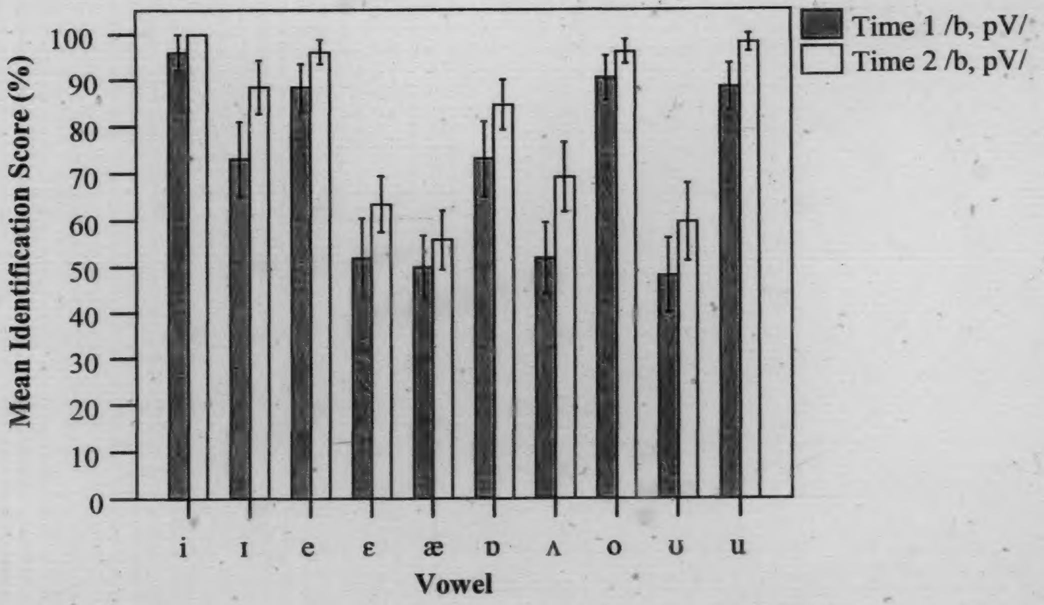


Figure 4.2. Pooled training groups' mean correct vowel identification scores on Voice 1 /b, pV/ stimuli, at Time 1 and 2. Error bars represent standard errors.

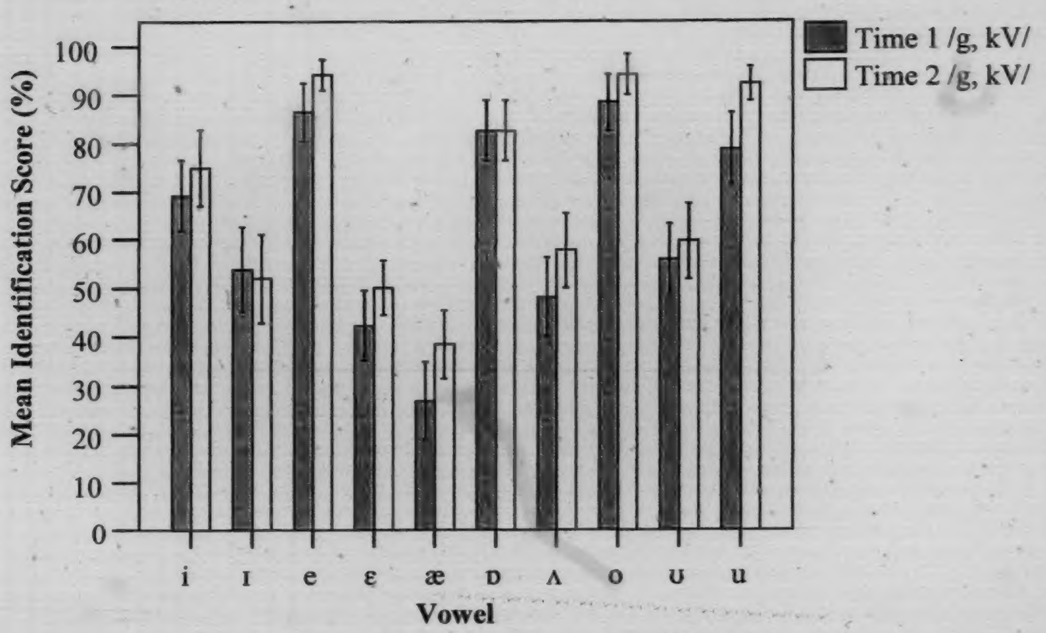


Figure 4.3. Pooled training groups' mean correct vowel identification scores on Voice 1 /g, kV/ stimuli, at Time 1 and 2. Error bars represent standard errors.

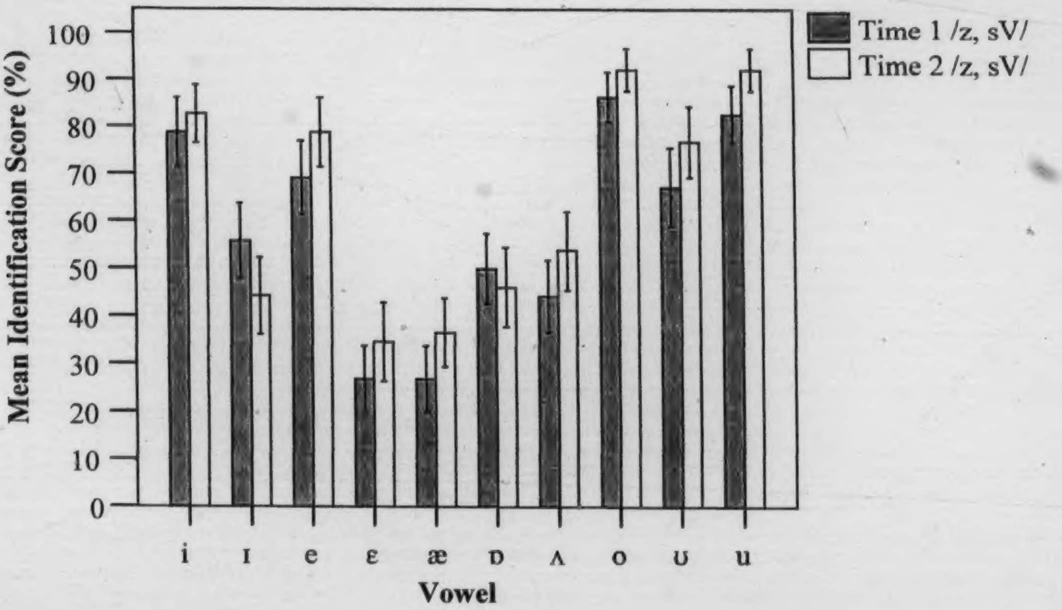


Figure 4.4. Pooled training groups' mean correct vowel identification scores on Voice 1 /z, sV/ stimuli, at Time 1 and 2. Error bars represent standard errors.

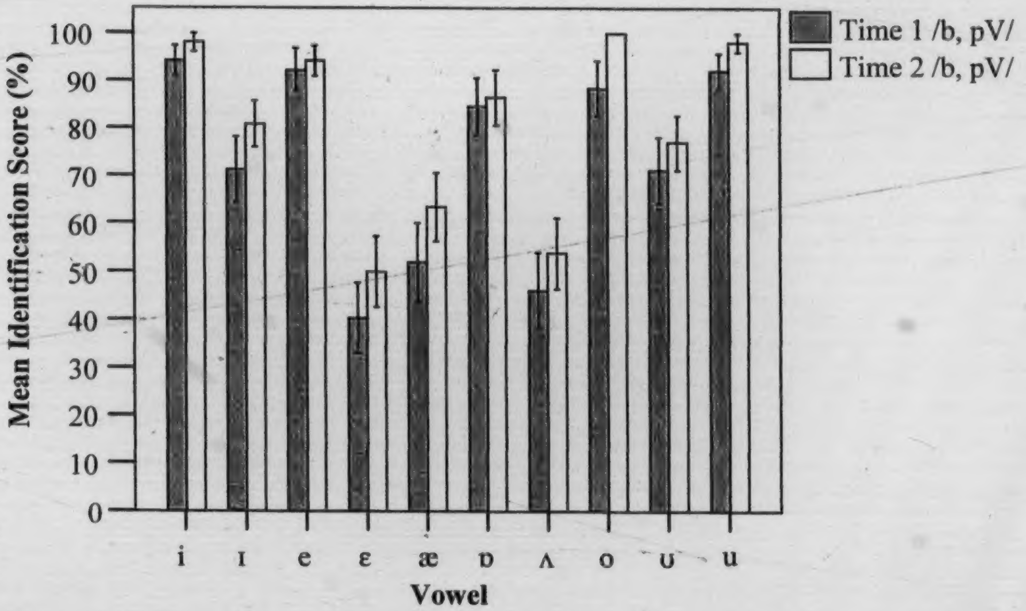


Figure 4.5. Pooled training groups' mean correct vowel identification scores on Voice 2 /b, pV/ stimuli, at Time 1 and 2. Error bars represent standard errors.

For the purposes of comparison with Voice 1 stimuli, responses to Voice 2 /b, pV/ stimuli are provided in Figure 4.5 above. As with Voice 1 /b, pV/ stimuli, mean identification rates for Voice 2 improved for every vowel. The Pearson correlation across Time 1 and 2 for mean vowel identification scores in response to the Voice 2 /b, pV/ stimuli is $r = .98$, again indicating that relative identification rates across vowels did not change.

Responses to Voice 1 and Voice 2 stimuli in the /b, pV/ context

Comparing differences in vowel identification rates on the basis of stimulus voice, a three-way partial repeated measures ANOVA was computed with Time (2 levels), Voice (2 levels) and Vowel (10 levels) as repeated measures and Training Group (3 levels) as between-subject factor. Again, because Mauchly's Test of Sphericity was significant for Vowel, corrected Huynh-Feldt measures are reported. A comparison of significant effects found using Huynh-Feldt versus uncorrected and multivariate Wilks' Lambda measures indicated similar results (See Table A5.2 in Appendix 5). Significant differences were found for Time [$F(1, 23) = 6.849, p < .01$] and Vowel [$F(8.054, 18.231) = 19.711, p < .01$], but not for Voice. A significant Voice x Vowel interaction was also found [$F(7.368, 169.466) = 2.712, p < .05$]. No other significant interactions between within-subject factors or with Training Group were found.

Although these results indicate no significant differences in overall improvement across stimulus voices in the same /b, pV/ context, post-hoc Tukey HSD tests found a significant difference in identification rates between Voice 2 /b, pV/ stimuli and Voice 1 /g, kV/ and /z, sV/ stimuli at Time 1 and Time 2.

The significant Voice x Vowel interaction suggests that L2 speakers' performance differed across vowels depending on what voice they heard as stimulus. The differences in mean identification scores for Voice 1 versus Voice 2 stimuli by Vowel are illustrated for Time 1 and Time 2 in Figures 4.6 to 4.7 respectively; results are pooled across training groups. Although post-hoc Tukey HSD and Bonferroni-adjusted *t*-tests comparing individual vowels by stimulus voice only indicate a significant difference in mean identification rates for English /ʊ/, the lack of significance differences in identification scores for other vowels may be due to conservative nature of multiple comparison tests, resulting in a lack of power. Comparisons of Vowel x Voice illustrated in Figures 4.6 and 4.7 indicate that mean identification scores for /ɛ/ and /ʌ/ are higher in response to Voice 1 stimuli than in response to Voice 2 stimuli, especially at Time 2.

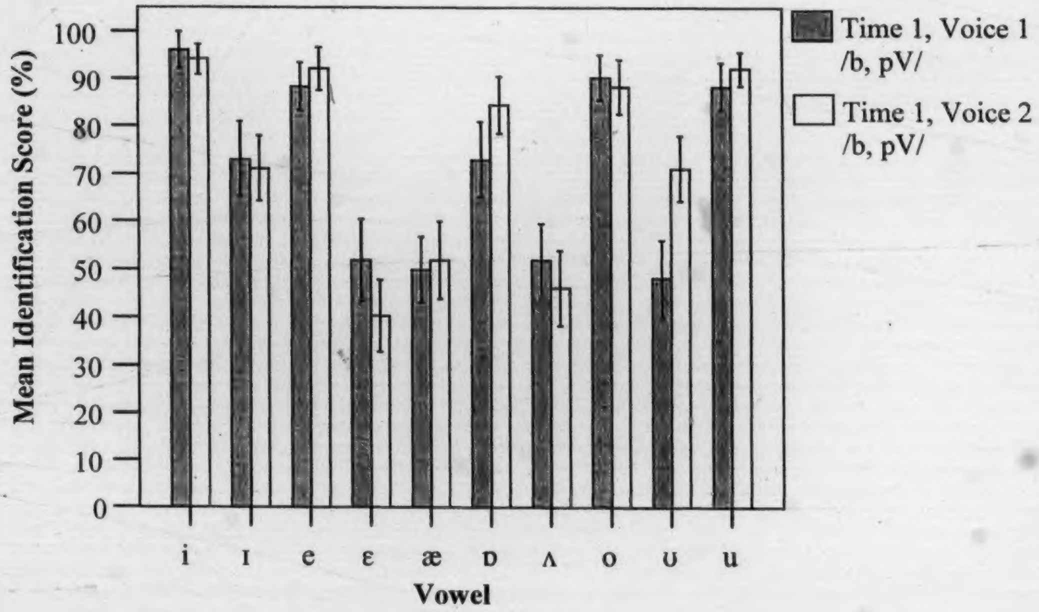


Figure 4.6. Pooled training groups' mean correct vowel identification scores on /b, pV/ stimuli by Voice 1 and 2 at Time 1. Error bars represent standard errors.

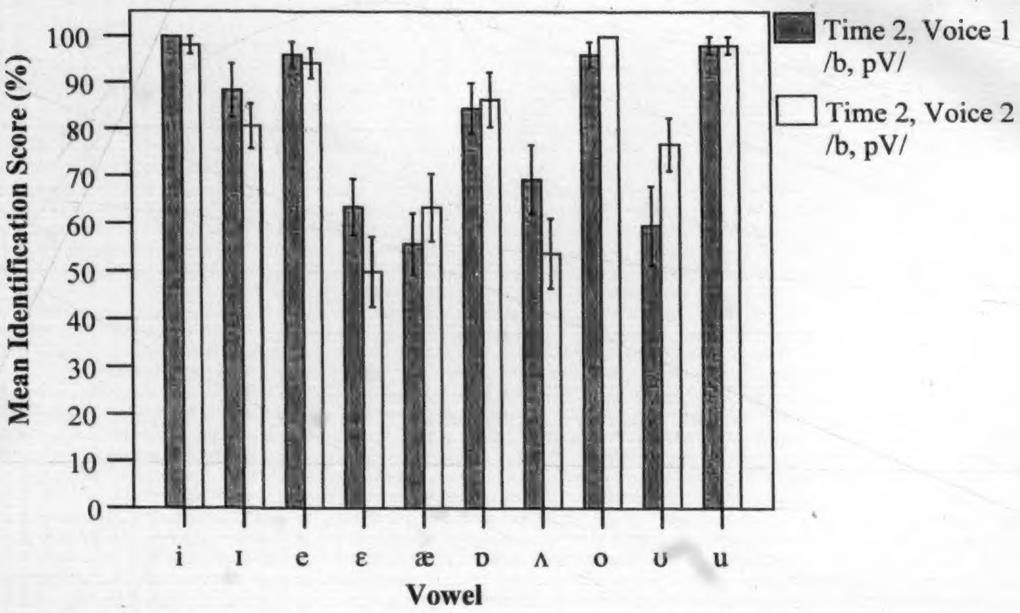


Figure 4.7. Pooled training groups' mean correct vowel identification scores on /b, pV/ stimuli by Voice 1 and 2 at Time 2. Error bars represent standard errors.

Summary of results for identification of English vowels on Generalization Test

All training groups demonstrated significant and roughly equal improvement in global English vowel identification rates from Time 1 to Time 2, although there is large variation in the relative performance across English vowel categories. Performance on some vowels reached a ceiling of nearly perfect identification, while performance on other vowels continued to lag far behind. Additionally, while improvement was significant for all consonantal contexts tested, improvement was greatest for the training context, /b, pV/. Finally, global improvement was equal in response to both test stimulus voices, but mean identification rates for a few categories differed depending on which voice was heard.

Lengthened Vowel and Natural Vowel tests on training items across time

The results of tests on training items for each training group over time are provided below. Recall that each participant was tested twice on his or her own training stimuli (Time 1 and Time 2) in addition to being tested on the opposing groups' tests of training stimuli at the end of training (Time 2). Mean correct identification rates for pooled vowels by CV context, speaker and time are provided in Table 4.3[14]. Because not all groups were tested on all items at Time 1, some Time cells in this table are not applicable, indicated as n/a.

Table 4.3. Mean % correct vowel identification scores and standard deviations by Lengthened Vowel vs Natural Vowel test , Training Group and Time.

	Long Test				Natural Test			
Group	LVT (n=11)	SVT (n=11)	DVT (n=4)	Mean	LVT (n=11)	SVT (n=11)	DVT (n=4)	Mean
Time 1	68.48	n/a	n/a	68.48	n/a	64.73	73.56	69.15
SD	*13.14*	*n/a*	*n/a*	*13.14*	*n/a*	*7.37*	*3.67*	*5.52*
Time 2	76.50	74.50	74.82	75.27	74.82	75.57	79.81	76.73
SD	*9.31*	*6.87*	*1.21*	*5.80*	*9.58*	*8.53*	*3.22*	*7.11*

A series of two-way repeated measures ANOVAs were computed for each training group (LVT, SVT, and DVT) with Time (2 levels) and Vowel (10 levels) as within-subject variables. For all three tests, Mauchly's Test of Sphericity was non-significant. Therefore, only uncorrected values are reported. A comparison of significant effects found using Huynh-Feldt versus uncorrected and multivariate Wilks' Lambda measures indicated similar results (See Table A5.3 in Appendix 5)

For the LVT group, significant differences were found in their performance on the lengthened vowel test for both Time [$F(1,10) = 16.251, p < .01$] and Vowel [$F(9,90) = 22.472, p < .01$]. No significant effect was found for the Time x Vowel interaction. Figure 4.8 illustrates mean identification scores for each vowel over Time. Mean vowel category identification scores across time were highly correlated $r = .99, p < .01$, indicating that the relative identification rates among vowel categories remained nearly the same over time.

---

[14] An additional comparison across groups is provided by looking at their mean identification scores on each training day. Although these are not comparable in terms of training items, average performance on each vowel suggest type of training yields similar results. By Training Day 2, all groups appeared to have learned the task and during their consecutive training days, mean identification scores across vowels increased in a linear fashion, highly correlated across groups (Mean Pearson correlations across groups, $r = .98$). Figure A5 in Appendix 5 illustrates these results.

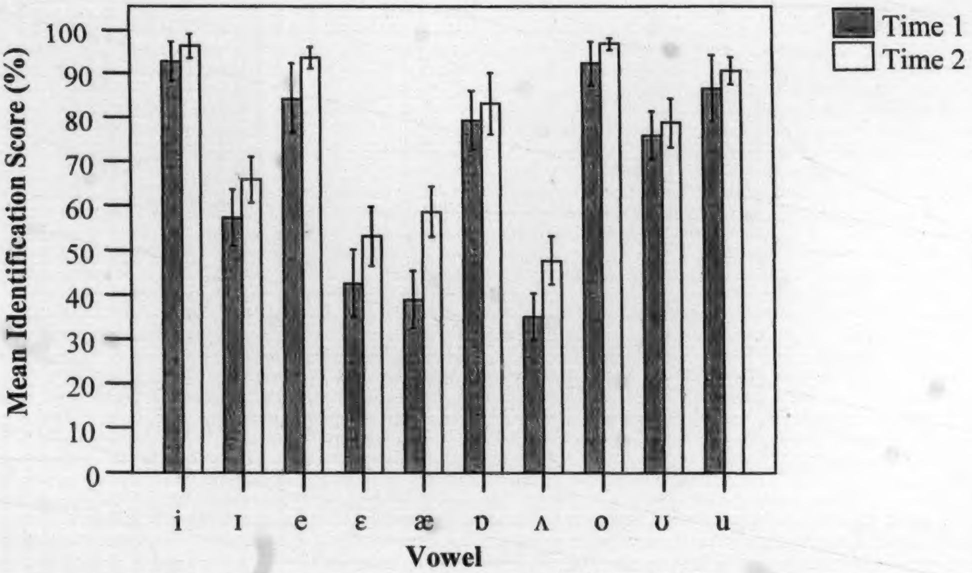


Figure 4.8. LVT group's mean correct vowel identification scores on lengthened-vowel training stimuli at Time 1 and Time 2. Error bars represent standard errors.

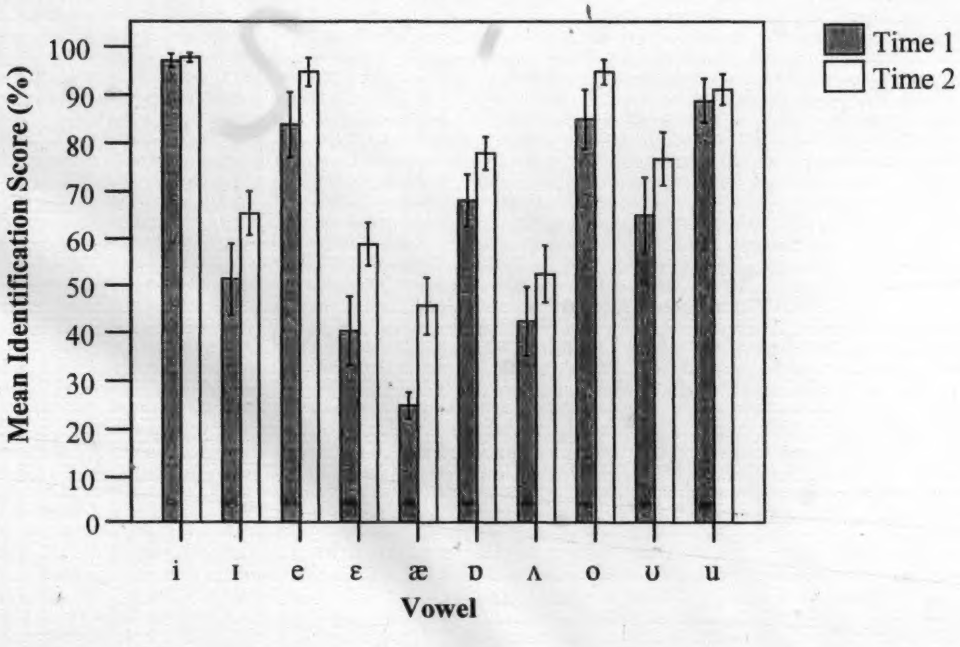


Figure 4.9. SVT group's mean correct vowel identification scores on natural vowel training stimuli at Time 1 and Time 2. Error bars represent standard errors.

For the SVT group, significant differences were found in their performance on the natural vowel test for both Time, [$F(1,10) = 12.601$, $< .01$] and Vowel, [$F(9,90) = 34.728$, $p < .01$]. No significant effect was found for the Time x Vowel interaction. Figure 4.9 illustrates mean identification scores for each vowel over time. Mean vowel identification scores across time were highly correlated, $r = .99$, $p < .01$, indicating that the relative identification rates among vowel categories remained nearly the same over time.

For the DVT Group, significant differences were found in their performance on the natural vowel test for both Time $F(1,3) = 52.083$, $p < .01$ and Vowel $F(9,27) = 4.591$, $p < .01$. Figure 4.10 illustrates mean identification scores for each vowel over time. Mean vowel identification scores across time were highly correlated, $r = .96$, $p < .01$, indicating that the relative identification rates among vowel categories remained nearly the same over time.

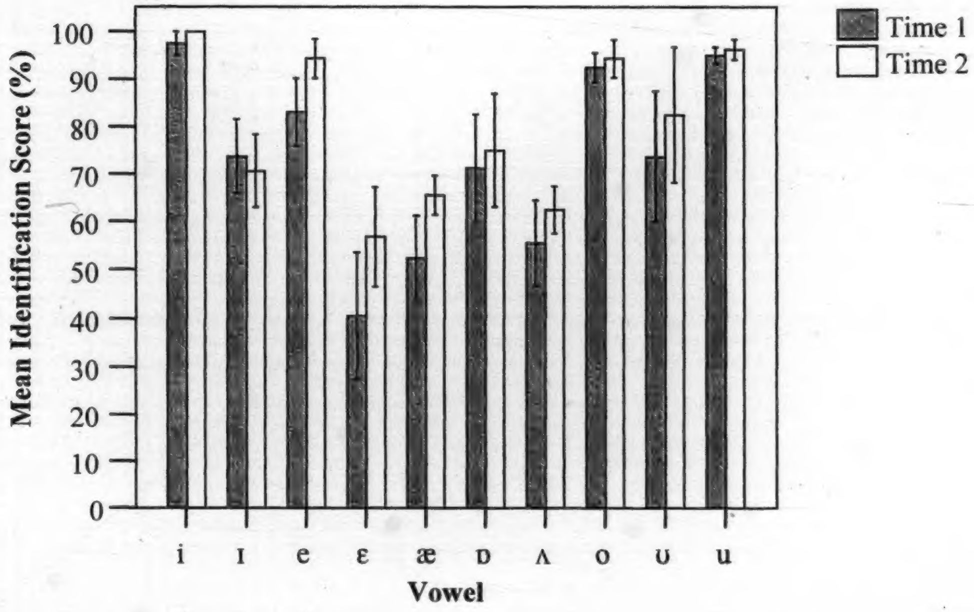


Figure 4.10. DVT group's mean correct vowel identification scores on natural vowel training stimuli at Time 1 and Time 2. Error bars represent standard errors. (Interpret with caution as n = 4).

Identification scores contrasting all learners on lengthened vs natural vowel tests

To test for differences in mean vowel identification scores between test types (Lengthened Vowel vs. Natural Vowel tests that all participants took at Time 2), a two-way partially repeated measures ANOVA was computed with Vowel Stimulus Length (2 levels) and Vowel (10 levels) as within-subject variables and Training Group as the between-subjects variable. Mauchly's Test of Sphericity was significant for Vowel, hence Huynh-Feldt corrected measures were used. Comparison of effects found using Huynh-Feldt versus other corrected, uncorrected and multivariate Wilks' Lambda measures indicate similar results (See Table A5.4 in Appendix 5). No significant effect was found for Vowel Stimulus Length. However, a significant difference was found for Vowel, [$F(7.512,172.777) = 42.831$, $p < .01$] as well as for the Vowel Stimulus Length x Vowel interaction, [$F(7.075,162.715) = 4.560$, $p < .01$]. The Partial Eta Squared statistics for Vowel and the Vowel Stimulus Length x Vowel interaction were .651 and .165 respectively, indicating a relatively weak effect size for the Vowel Stimulus Length x Vowel interaction. No significant differences were found across training groups.

Differences in mean identification scores on the Natural Vowel test compared to identification scores on the Lengthened Vowel test are illustrated in Figures 4.11; results are pooled across training groups. The Pearson correlation between mean vowel identification rates on natural versus lengthened vowels was $r = .97$, indicating that relative performance across vowel categories is similar, regardless of whether the vowel stimuli is natural or has been artificially lengthened.

Post-hoc Tukey tests on all data pooled across groups indicated a significant difference between lengthened and natural vowel conditions for only one vowel, /æ/, which was more accurately identified in the lengthened vowel condition. Figure 4.11 suggests that there is also a possible difference in performance on /ʌ/, which appears to be more accurately identified in the natural vowel condition. Further discussion of these differences will be presented in the next chapter.

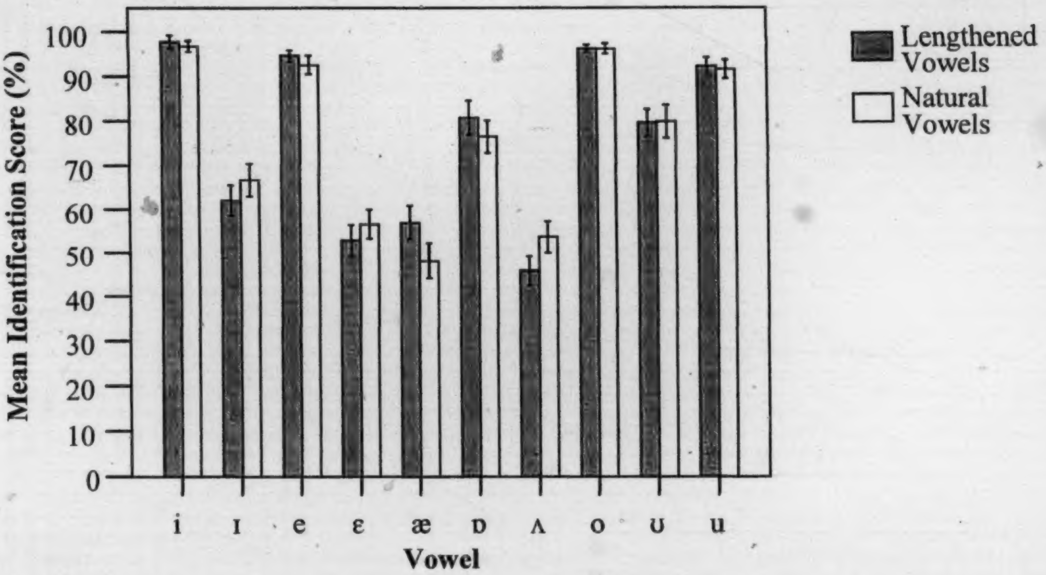


Figure 4.11. Pooled training groups' mean correct identification scores on natural versus lengthened vowel stimuli at Time 2. Error bars represent standard errors.

Since no effect for Training Group (particularly between SVT and DVT) was found, this seems to indicate that there is no clear advantage to training on a select set of English vowel stimuli (i.e., those that are least Mandarin-like). Of course, this conclusion is limited by the sample sizes in each training condition. Recall that in the SVT group the sample size was eleven, while in the DVT group it was only four. To determine if there was any measurable difference on the mean identification rates for the less Mandarin-like versus more Mandarin-like items, I conducted a follow-up two-way partially repeated measures ANOVA, on the Time 2 Lengthened and Natural Vowel test data from all groups to assess any differences in identification scores on the basis of these token subclasses (i.e., those deemed more Mandarin-like and those deemed less Mandarin-like). Mean identification scores for each subclass and test-type are provided in Table 4.4. Test stimuli subclasses (more or less Mandarin-like) and training stimuli (lengthened vowels and natural vowels) served as within-subject measures, while Training Group served as a between-subjects measure. Results indicate that regardless of test stimuli (natural or lengthened vowel), performance on less Mandarin-like stimuli was significantly better [$F(1,597) = 7.820$, $p < .01$] than on the more Mandarin-like stimuli. However, the mean difference was only 3.23% on the natural items and 2.94% on the lengthened vowel items and the effect size for subclasses of English vowels was very small [Partial Eta Squared = .013].

Table 4.4. Mean % correct identification scores and standard deviations pooled across all subjects for each subclass of tokens.

	Less Mandarin-like	More Mandarin-like	Difference
Natural Vowel Test % correct	77.52	74.29	3.23
SD	*8.64*	*8.59*	
Lengthened Vowel Test % correct	77.65	74.71	2.94
SD	*7.34*	*8.13*	

Summary of results for identification of natural versus lengthened vowel tests. All training groups demonstrated significant improvement in mean English vowel identification rates on their group-specific training tests from Time 1 to Time 2, although as with the Generalization test, there was large variation in relative identification rates across English vowel categories. Performance on some vowels reached a ceiling of nearly perfect identification, while performance on other vowels continued to lag far behind. Comparing performance at Time 2 on the lengthened vowel versus natural vowel tests, no significant difference in the main effect of test type was found, although identification of two vowels, /æ/ and /ʌ/, seemed to vary depending on test type. Finally, comparing performance on those vowel stimuli times that were deemed to be less Mandarin-like with those that were deemed to be more Mandarin-like, a significant but small effect in favour of the less Mandarin-like stimuli was found.

Delayed post-test, Generalization Test

To determine if the effect of English vowel identification training would persist over time without ongoing training, 18 participants (9 from LVT and 9 from SVT) agreed to participate in a test of retention approximately 1 month (*M* = 3.9 days, Range, 26-34) after finishing the last phase of training and testing. Their delayed post-test identification scores on the Generalization test were compared with their performance on the Time 2 tests immediately after training.

Mean correct identification scores and standard deviations on the Generalization test over time are provided in Table 4.5 below. Although I was primarily interested in testing whether performance decreased between Time 2 and the delayed post-test, to allow comparison with initial performance, Time 1 means and standard deviations are also provided.

Table 4.5. Mean % correct vowel identification scores and standard deviations on the Generalization test by CV context, and stimulus Voice at Time 1, Time 2 and Delayed post-test.

	/b, pV/ Voice 1		/g, kV/ Voice 1		/z, sV/ Voice 1		/b, pV/ Voice 2		Mean across groups and CVs
Group	LVT (n=9)	SVT (n=9)	LVT (n=9)	SVT (n=9)	LVT (n=9)	SVT (n=9)	LVT (n=9)	SVT (n=9)	
Time 1	74.44	67.78	66.67	62.78	61.11	58.89	78.89	68.33	67.36
SD	*14.46*	*20.48*	*16.20*	*21.95*	*13.18*	*19.65*	*9.61*	*16.96*	*16.56*
Time 2	81.67	85.00	76.67	66.67	68.33	64.44	82.22	80.00	76.81
SD	*12.75*	*10.31*	*14.36*	*15.21*	*16.77*	*13.57*	*11.49*	*11.46*	*13.24*
Delayed Post-test	80.56	79.44	74.44	70.56	71.67	64.44	80.56	78.33	74.51
SD	*13.56*	*8.46*	*12.61*	*12.86*	*13.69*	*14.24*	*12.36*	*11.99*	*12.47*

Differences in performance comparing /b, pV/-/g, kV/-/z, sV/ contexts for Voice 1 are treated separately from comparisons of differences in performance in response to each voice in the /b,pV/ context; recall that Voice 2 stimuli only included the /b,pV/ context.

Responses to Voice 1 stimuli in /b, pV/, /g, kV/ and /z, sV/ contexts

A three-way partially repeated measures ANOVA was computed with Time (2 levels, Time 2 and the delayed post-test) Consonant (Voice 1 only, 3 levels) and Vowel (10 levels) as repeated measures, and Training Group (LVT, SVT) as a between-subject factor. Mauchly's Test of Sphericity was significant for Vowel, therefore, corrected

Huynh-Feldt measures are reported. Comparison of effects found using Huynh-Feldt versus other corrected, uncorrected and multivariate Wilks' Lambda measures indicate similar results (See Table A5.5 in Appendix 5). Significant differences were found for Consonant [$F(1.986, 31.780) = 22.079, p < .01$] and Vowel [$F(5.136, 82.176) = 17.475, p < .01$]. A significant Consonant x Vowel interaction was also detected [$F(11.988, 191.808) = 4.502, p < .01$]. No significant effect of Time was found. Post-hoc Tukey HSD tests on the effect of Consonant showed a significant difference in correct identification rates at Time 2 for Voice 1 /b, pV/-/z, sV/ as well as /b, pV/-/g, kV/ contrasts; at the delayed post-test, significant differences in identification rates were found for Voice 1 /b, pV/-/z, sV/ but not /b, pV/-/g, kV/ contrasts. In all cases, performance on /b, pV/ was better than the opposing contrast.

Differences in mean identification scores over time for each Vowel x Consonant combination for the Voice 1 stimuli are illustrated in Figures 4.12 to 4.14; results are pooled across training groups. The Pearson correlation across Time 2 and the delayed post-test for mean vowel identification scores in response to the Voice 1 /b, pV/ stimuli is $r = .98$; /g, kV/ stimuli, $r = .96$; and /z, sV/ stimuli, $r = .89$. This indicates that those vowels that had relatively weaker identification scores at Time 2, were still relatively the same at the delayed post-test.

While there was no significant interaction between Time and Vowel category, Figures 4.12 to 4.14 suggest some variation in performance on particular vowels from Time 2 to the delayed post-test. However, this variation is not consistent across CV contexts.

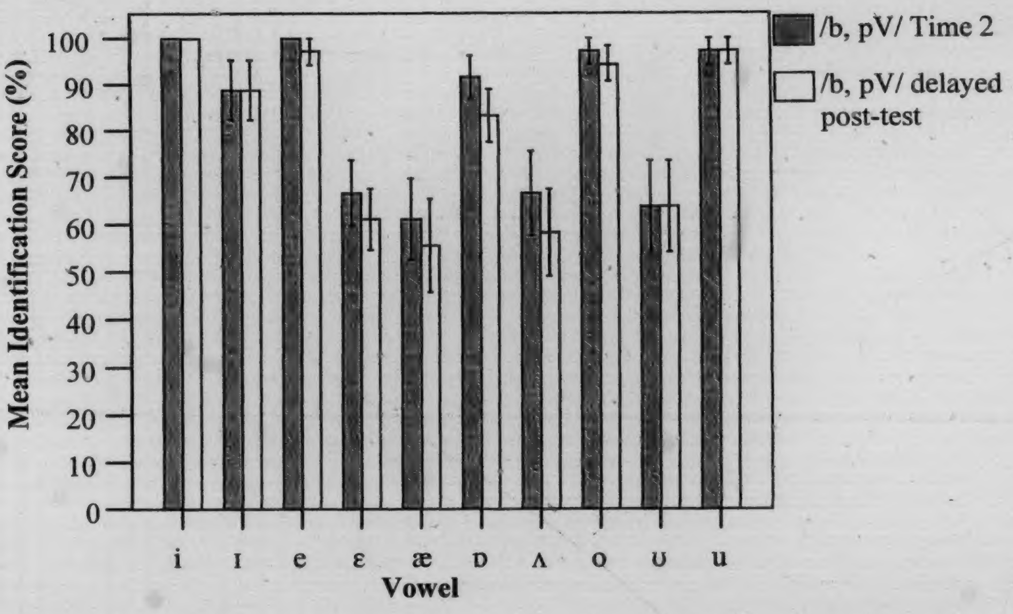


Figure 4.12. Pooled training groups' mean correct vowel identification scores on Voice 1 /b, pV/ stimuli, at Time 2 and delayed post-test. Error bars represent standard errors.

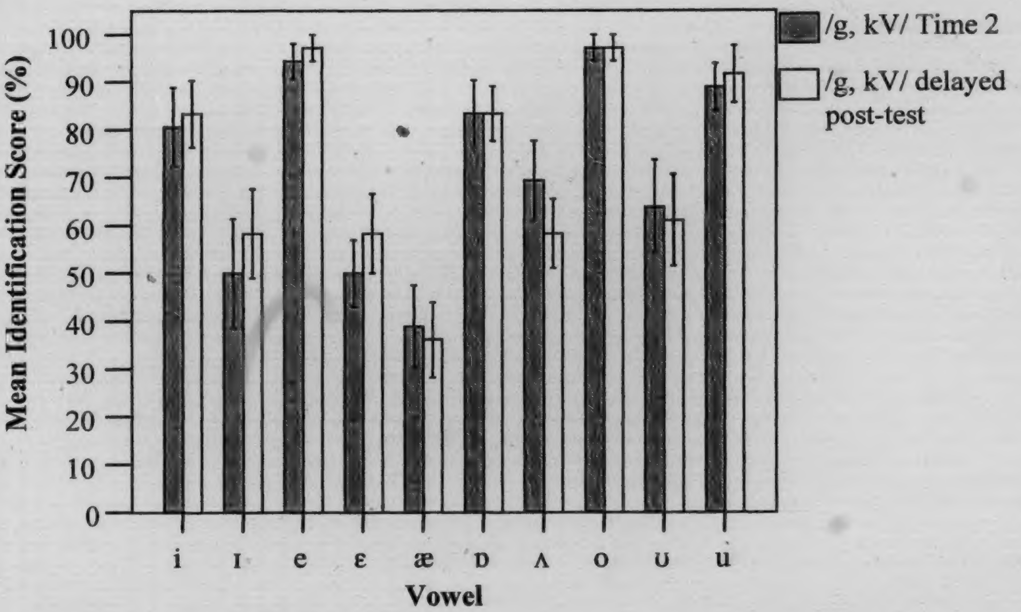


Figure 4.13. Pooled training groups' mean correct vowel identification scores on Voice 1 /g, kV/ stimuli, at Time 2 and delayed post-test. Error bars represent standard errors.

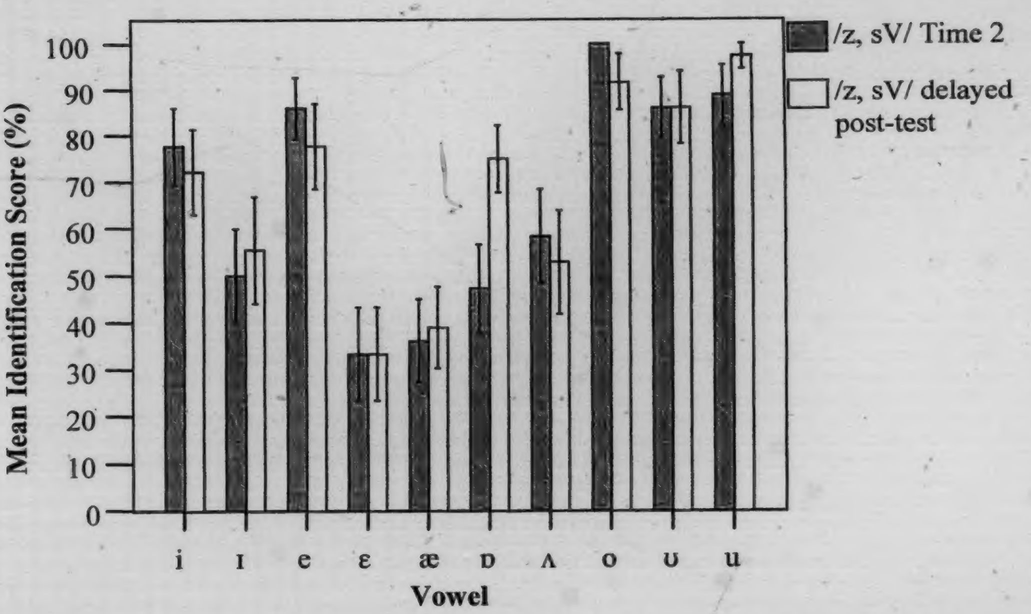


Figure 4.14. Pooled training groups' mean correct vowel identification scores on Voice 1 /z, sV/ stimuli, at Time 2 and delayed post-test. Error bars represent standard errors.

For the purposes of comparison with Voice 1 stimuli, responses to Voice 2 /b, pV/ stimuli are provided in Figure 4.15. As with Voice 1 /b, pV/ stimuli, some variation between across vowels at Time 2 compared to the delayed post-test is evident, however the general global pattern of no change is apparent. The Pearson correlation across Time 2 and the delayed post-test for mean vowel identification scores in response to the Voice 2 /b, pV/ stimuli is *r* = .97, again indicating that relative identification rates across vowels did not change appreciably.

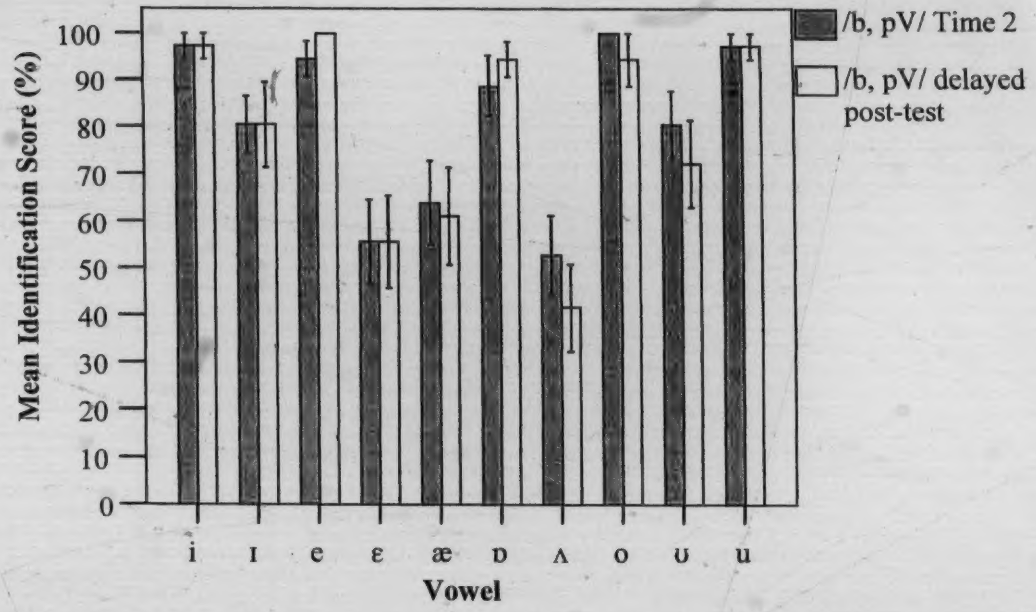


Figure 4.15. Pooled training groups' mean correct vowel identification scores on Voice 2 /b, pV/ stimuli, at Time 2 and delayed post-test. Error bars represent standard errors.

Responses to Voice 1 and Voice 2 stimuli in the /b, pV/ context

Comparing differences in vowel identification rates on the basis of stimulus voice, a three-way partially repeated measures ANOVA was computed with Time (2 levels), Voice (2 levels) and Vowel (10 levels) as repeated measures and Training Group (3 levels) as the between-subjects factor. Again, because Mauchly's Test of Sphericity was significant for Vowel, corrected Huynh-Feldt measures are reported. Comparison of significant effects found using Huynh-Feldt versus other corrected, uncorrected and multivariate Wilks' Lambda measures indicate similar results (See Table A5.6 in Appendix 5). Significant differences were found for Vowel [*F*(6.250, 99.998) = 15.652, p < .01], but not for Voice or Time. A significant Voice x Vowel interaction [*F*(6.162,98.591) = 2.183, p = .049] was also found.

Although these results indicate no significant difference across stimulus voices in the same /b, pV/ context, post-hoc Tukey tests indicated a significant difference in identification rates between Voice 2 /b, pV/ stimuli and Voice 1 /z, sV/ stimuli at Time 2 as well as at the delayed Post-test, with performance on Voice 2 /b, pV/ stimuli being stronger. No significant differences were found between Voice 2 /b, pV/ and Voice 1 /g, kV/ contrasts at either Time 2 or the delayed post-test.

Differences in mean identification scores for Voice 1 versus Voice 2 stimuli by Vowel are illustrated for Time 2 and the delayed post-test in Figures 4.16 and 4.17 respectively; results are pooled across training groups. As was discovered for Time 1 and Time 2 data previously, participants' mean identification scores on some vowels are better for Voice 1 than Voice 2 and vice versa.

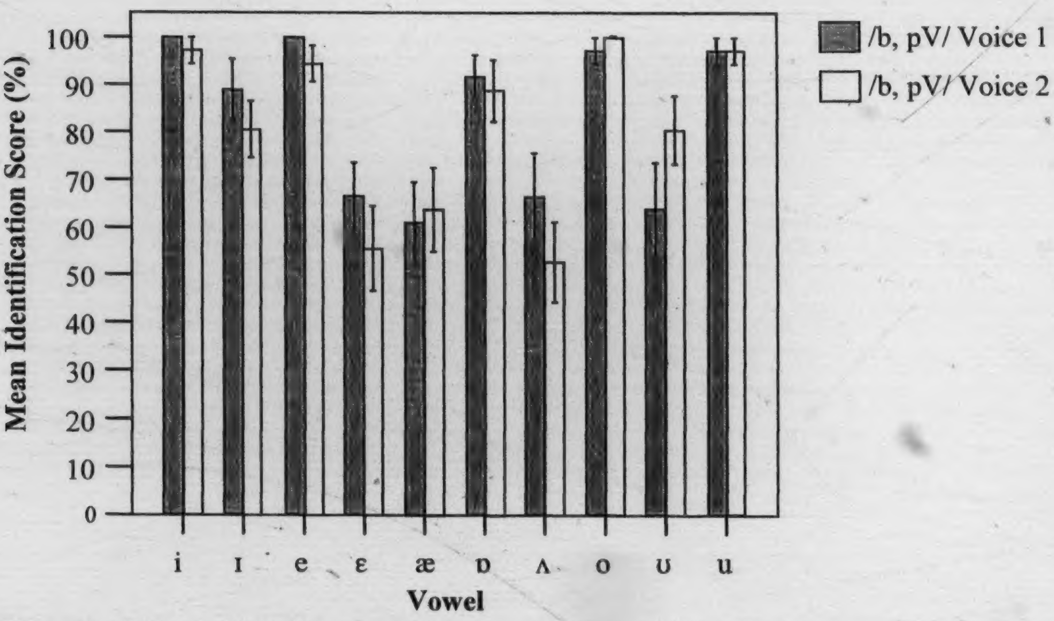


Figure 4.16. Pooled training groups' mean correct vowel identification scores on /b, pV/ stimuli by Voice 1 and 2 at Time 2. Error bars represent standard errors.

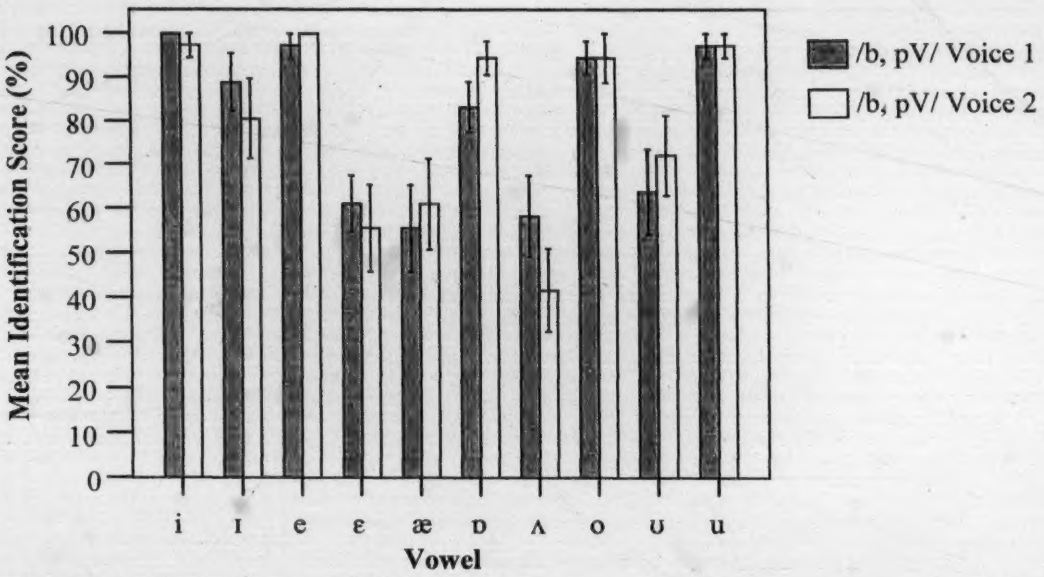


Figure 4.17. Pooled training groups' mean correct vowel identification scores on /b, pV/ stimuli by Voice 1 and 2 at delayed post-test. Error bars represent standard errors.

Natural Vowel Test

Mean scores comparing the Time 2 test with the Delayed post-test are shown in Table 4.6. below.

Table 4.6. Mean % correct identification scores and standard deviations on the Natural Vowel test by Training Group and Time

	Natural Vowel Test		
Training Group	LVT (n=9)	SVT (n=9)	Average
Time 2	76.94	75.97	76.46
SD	*7.94*	*9.36*	*8.65*
Delayed post-test	77.78	75.44	76.61
SD	*7.52*	*7.34*	*7.43*

To test for differences in vowel identification rates from the Time 2 natural vowel training stimuli test to the delayed post-test, a two-way partially repeated measures ANOVA was computed with Time (2 levels) and Vowel (10 levels) as within-subject variables and Training Group as the between-subjects factor. Mauchly's Test of

Sphericity was significant. Therefore, Huynh-Feldt values are reported (Table A5.7 in Appendix 5 provides a comparison of uncorrected, corrected and multivariate analyses results).

Significant differences were found for Vowel [$F(5.556,88.891) = 45.588$, $p < .01$], but not Time or Training Group. A significant effect was also found for the Time x Vowel interaction [$F(7.503,120.043) = 2.991$, $p < .01$]. However, the effect size for this interaction (Partial Eta Squared = .147) was small, while the effect size of Vowel (Partial Eta Squared = .740) is large. Figure 4.18 illustrates mean identification scores pooled across Training group for each vowel at Test 2 and the delayed post-test. Mean vowel identification scores across time were highly correlated [$r = .98$, $p < .01$], indicating that relative performance on each vowel category remained largely unchanged between Time 2 and the delayed post-test.

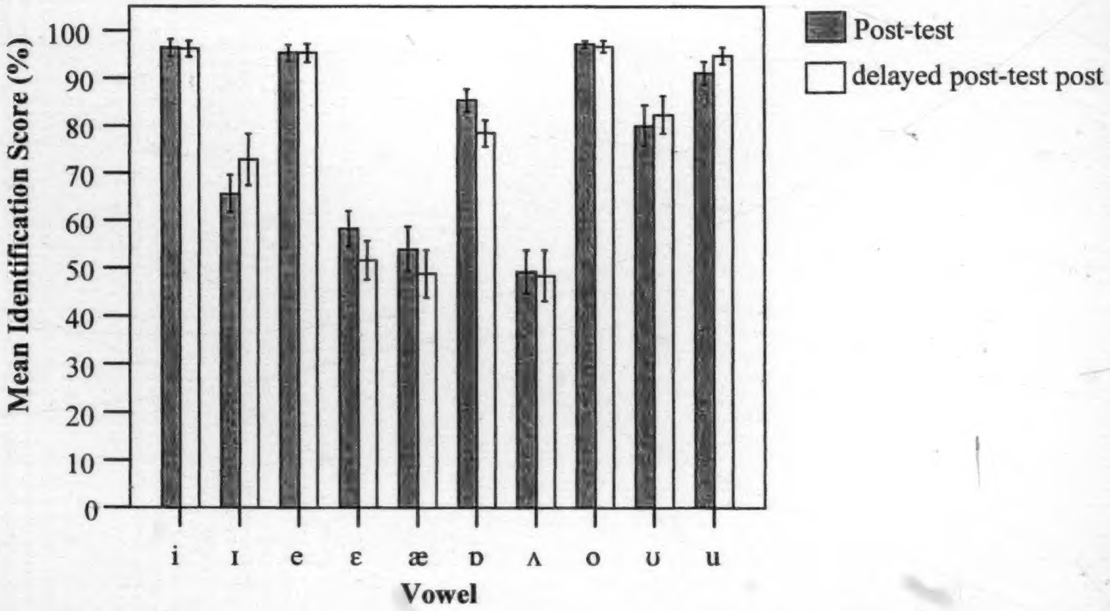


Figure 4.18. Pooled groups' mean correct vowel identification scores on natural vowel training stimuli at Time 2 and delayed post-test. Error bars represent standard errors.

Summary of results comparing Time 2 with delayed post-test.

Those 18 participants who returned for the delayed post-test 4 weeks after training showed no significant increase or decrease in their ability to identify English vowels in either the Generalization test or Natural Vowel test. In addition, at the delayed Generalization post-test, they continued to perform better on /b, pV/ items than on the other /g, kV/ and /z, sV/. Finally, there was still no indication that stimuli produced by one voice in the Generalization test resulted in overall stronger identification rates than stimuli presented with a second voice. However, as before, some differences were indicated for specific vowels produced by each of the voices.

4.3.2. Vowel production test

The results of the vowel production test on the same CV syllables used in the identification test of generalization are presented below. Recall that for production, two repetitions of each CV syllable were elicited. A two-way repeated measures ANOVA was computed with Time (2 levels) and Repetition (2 levels) as repeated measures to determine if any difference existed between performance on each production repetition. Although significant improvement in production was found from Time 1 to Time 2 [$F(1,1759) = 10.412, p < .01$], differences between Repetitions were not significant, nor was there any significant Time x Repetition interaction. The Mean Square for repetitions was .119, indicating very little variance. This finding allowed for the pooling of correct vowel scores across repetitions. Mean correct identification rates for pooled vowels by CV context, speaker and time are provided in Table 4.7.[15] Mean correct production rates by vowel category are provided in later figures. Differences in performance comparing /b, pV/-/g, kV/-/z, sV/ contexts for Voice 1 are treated separately from comparisons of differences in performance in response to each voice in the /b,pV/ context; recall that Voice 2 stimuli only included the /b,pV/ context.

---

[15] Because the production data in this experiment was tested against the English Model that was trained using only /b, pV/ contexts, including duration as a variable may not accurately reflect differences across /b, pV/-/g, kV/-/z, sV/ contexts. Therefore, I report results here in terms of the English Model classifications that exclude duration as a variable so as not to skew results in favor of the /b, pV/ contexts over the other contexts. By excluding duration, results reflect spectral improvement only. The same tests were conducted using the English Model including duration and the global pattern of results were not substantially different than when I excluded duration. A version of the results reported in this section, but including vowel duration as a variable, are provided in Table A6.1 –A6.6 in Appendix 6.

Table 4.7. Mean % correct vowel production recognition scores and standard deviations on the production test as recognized by the English Model. Results are provided by CV context, stimulus Voice, Training Group and Time. Vowel duration was excluded as a factor in the English CV pattern recognition model.

	/b, pV/ Voice 1		/g, kV/ Voice 1		/z, sV/ Voice 1	
Group	LVT (n=11)	SVT (n=11)	LVT (n=11)	SVT (n=11)	LVT (n=11)	SVT (n=11)
Time 1	70.00	65.23	72.05	68.64	67.27	62.95
SD	*5.24*	*11.59*	*10.11*	*12.57*	*11.15*	*12.14*
Time 2	75.68	71.36	71.14	67.95	71.14	62.73
SD	*9.43*	*11.75*	*11.03*	*13.50*	*10.63*	*13.71*

	/b, pV/ Voice 2		Average across groups and CVs
Group	LVT (n=11)	SVT (n=11)	
Time 1	67.73	70.23	68.01
SD	*7.28*	*5.06*	*9.39*
Time 2	72.27	74.77	70.88
SD	*9.45*	*10.46*	*11.25*

Responses to Voice 1 stimuli in /b, pV/, /g, kV/ and /z, sV/ contexts

A three-way partially repeated measures ANOVA was computed with Time (2 levels) Consonant (Voice 1 only, 3 levels) and Vowel (10 levels) as repeated measures, and Training Group (LVT, SVT) as the between-subjects factor. Because Mauchly's Test of Sphericity was significant for Vowel, corrected Huynh-Feldt measures are reported. Comparison of effects found using Huynh-Feldt versus other corrected, uncorrected and multivariate Wilks' Lambda measures indicate similar results (See Table A5.8 in Appendix 5). Significant differences were found for Vowel [$F(7.154,143.080) = 21.569, p < .01$], but no other significant main effects either within or between groups were detected. Significant Consonant x Vowel [$F(15.371, 307.424) = 9.167, p < .01$] and Time x Consonant [$F(1.910,38.194) = 4.291, p < .01$] interactions were also found.

Post-hoc HSD Tukey tests on the effect of Consonant indicated Voice 1 productions of vowels in /g, kV/ syllables were recognized by the English Model as being significantly more accurate than vowels produced in /z, sV/ syllables at Time 1; there were no significant differences in recognition scores for vowels in /b, pV/-/z, sV/ or /b, pV/-/g, kV/ contrasts at Time 1; at Time 2, vowels in /b, pV/ syllables were recognized as the intended vowel significantly more often than vowels in /z, sV/ contexts; there were no significant differences in recognition scores for vowels in /b, pV/-/g, kV/ or /g, kV/-/z, sV/ contrasts.

To further explore possible sources of the significant Time x Consonant interaction, additional simple effects two-way repeated measures ANOVAs were conducted to measure mean differences in performance over time for each of the three CV contexts independently. For each CV context, Time (2 levels) and Vowel (10 levels) served as within group factors. Mauchly's test of sphericity was significant for Vowel, therefore corrected Huynh-Feldt measures are reported. A comparison of effects found using Huynh-Feldt versus other corrected, uncorrected and multivariate Wilks' Lambda measures indicate similar results (See Tables A5.9 through A5.11 in Appendix 5). For the /b, pV/ context, a significant difference was found for Time [$F(1,20) = 7.660$, $p = .012$] as well as for Vowel [$F(6.453,129.051) = 28.682$, $p < .01$]. For the /g, kV/ and /z, sV/ contexts, Time was not significant, although Vowel was for both /g, kV/, [$F(5.635,112.708) = 12.393$, $p < .01$] and /z, sV/ [$F(6.640,132.808) = 12.283$, $p < .01$]. Differences in mean identification scores over time for each Consonant x Vowel combination for the Voice 1 stimuli are illustrated in Figures 4.19 to 4.21; results are pooled across training groups. The Pearson correlation across Time 1 and 2 for mean vowel identification scores in response to the Voice 1 /b, pV/ stimuli is $r = .96$; /g, kV/ stimuli, $r = .95$; and /z, sV/ stimuli, $r = .92$, indicating that the relative accuracy on each English vowel category remained quite stable over time. However, there are some differences in the degree to which individual sounds improved. Figure 4.19 shows that the largest improvement in production in the /b, pV/ context occurred in /ɪ/, /æ/ and /ɒ/ and /ʌ/, with little if any improvement in other vowels. However, with the exception of /u/, other vowels were already recognized in production at closer to ceiling levels of accuracy at Time 1.

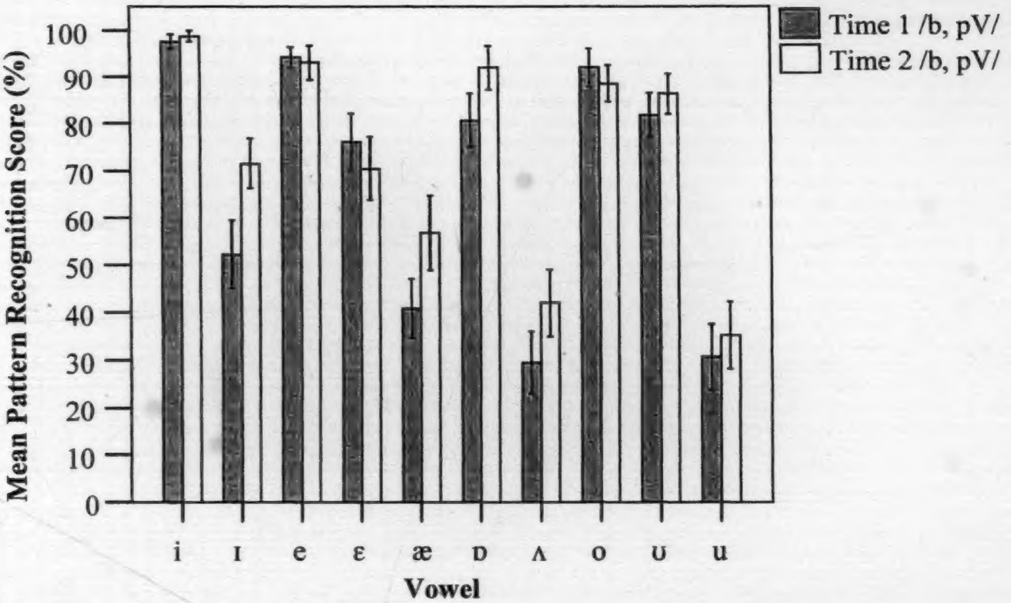


Figure 4.19. Pooled groups' mean percent correct production scores over time in response to /b, pV/ stimuli produced by Voice 1. Error bars represent standard errors.

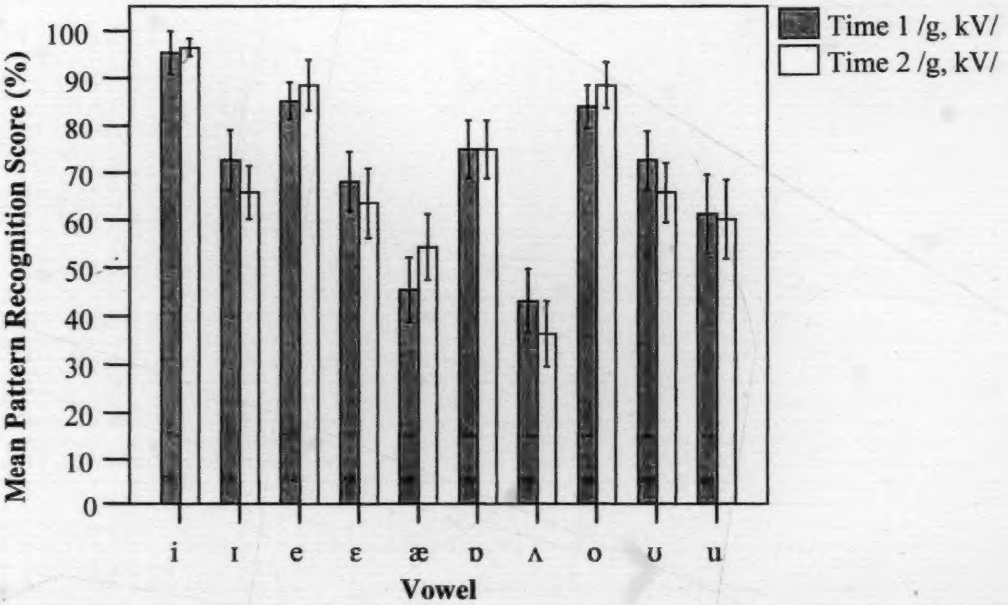


Figure 4.20. Pooled groups' mean percent correct production scores over time in response to /g, kV/ stimuli produced by Voice 1. Error bars represent standard errors.

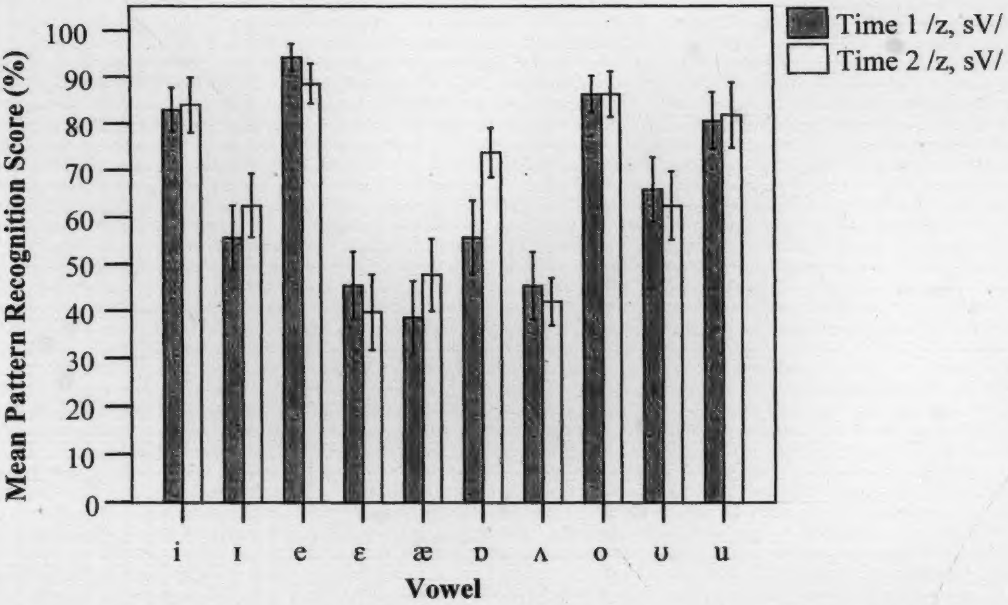


Figure 4.21. Pooled groups' mean percent correct production scores over time in response to /z, sV/ stimuli produced by Voice 1. Error bars represent standard errors.

For the purposes of comparison with Voice 1 /b, pV/ stimuli, responses to Voice 2 /b, pV/ stimuli are provided in Figure 4.22. The Pearson correlation across Time 1 and 2 for mean vowel identification scores in response to the Voice 2 /b, pV/ stimuli is *r* = .98. Clearest improvement is evident for /æ/ and to a lesser extent, for /ɒ/ and /u/.

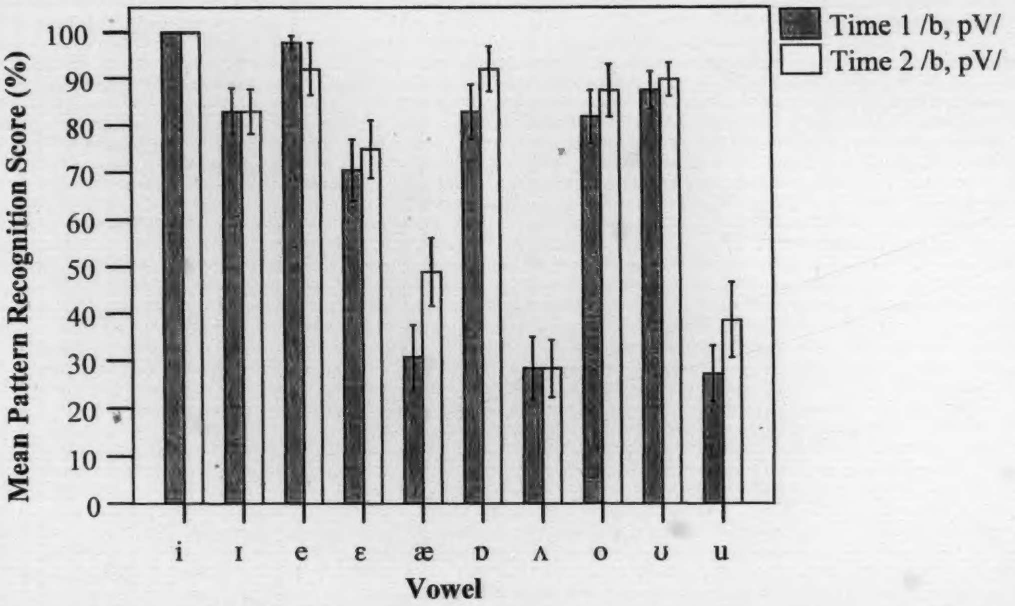


Figure 4.22. Pooled groups' mean percent correct production scores over time in response to /b, pV/ stimuli produced by Voice 2. Error bars represent standard errors.

Responses to Voice 1 and Voice 2 stimuli in the /b, pV/ context

Comparing differences in vowel identification rates on the basis of stimulus voice, a three-way partial repeated measures ANOVA was computed with Time (2 levels), Voice (2 levels) and Vowel (10 levels) as repeated measures and Training Group (2 levels) as the between-subjects factor. Again, because Mauchly's Test of Sphericity was significant for Vowel, corrected Huynh-Feldt measures are reported. Comparison of significant effects found using Huynh-Feldt versus uncorrected and multivariate Wilks' Lambda measures indicate similar results (See Table A5.12 in Appendix 5). Significant differences were found for Time [$F(1, 20) = 9.463, p < .01$] and Vowel [$F(5.804,116.076) = 39.278, p < .01$], but not for Voice. A significant Voice x Vowel interaction was also found [$F(6.233,124.660) = 4.515, p < .01$]. No other significant interactions between within-subject factors or with Training Group were found.

Post-hoc Tukey HSD tests comparing mean production recognition scores for each vowel produced in response to each stimulus voice found that at Time 1 productions of English /ɪ/ in response to the Voice 2 stimulus were more accurate than productions of

English /ɪ/ in response to the Voice 1 stimulus. Bonferroni adjusted *t*-tests indicated the same result. While no other statistically significant differences were found, other potential differences in mean production scores in response to Voice 1 versus Voice 2 stimuli are suggested by raw mean differences as illustrated in Figures 4.23 to 4.24 for Time 1 and Time 2 respectively. For example, at Time 1, productions in response to Voice 1 /æ/ and /o/ appear to be more accurate than productions in response to those vowels produced by Voice 2. At Time 2, productions in response to Voice 1 /ʌ/ appear to be much more accurate than responses to the same vowel by produced by Voice 2. Finally, productions in response to Voice 2 /ɪ/ appear to be more accurate than those in response to Voice 1 /ɪ/.

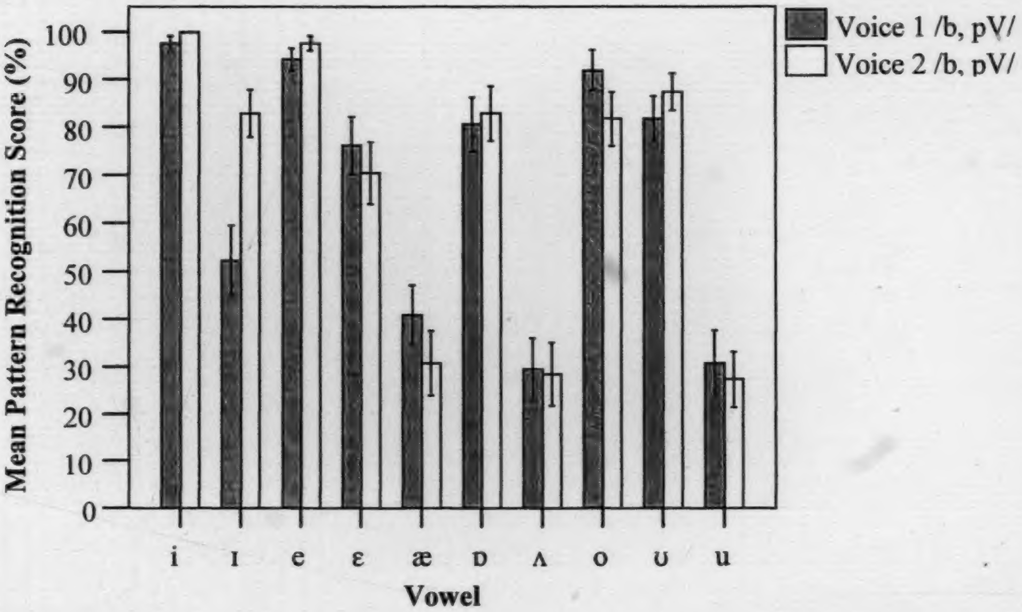


Figure 4.23. Pooled groups' mean percent correct production scores at Time 1 in response to /b, pV/ stimuli produced by Voice 1 and Voice 2. Error bars represent standard errors.

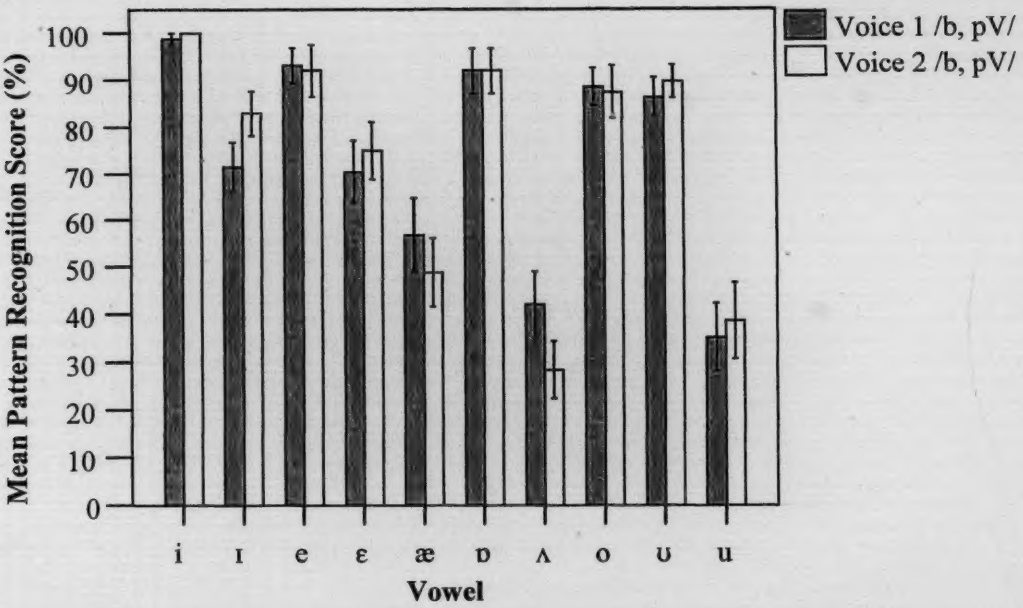


Figure 4.24. Pooled groups' mean percent correct production scores at Time 2 in response to /b, pV/ stimuli produced by Voice 1 and Voice 2. Error bars represent standard errors.

Summary of results for production

Participants' global production ability on the ten English vowels for which they received training showed significant improvement in the /b, pV/ context in which they were trained. However, improvement was not detected in the non-training contexts, /g, kV/ and /z, sV/. While the overall mean production accuracy rates of speakers' in response to Voice 1 versus Voice 2 stimuli were not significantly different, there appeared to be some differences in response to each stimulus Voice for a few vowels.

4.3.3. Relationship between identification and production results

To provide a general comparison of the learners' identification and production results, Figures 4.25 – 4.28, below, illustrate mean differences in the learners' (n = 22) Generalization test identification scores for /b, pV/ stimuli, compared with the same learners' production recognition scores for /b, pV/ stimuli. Recall that the same /b, pV/ stimuli recordings were used for identification and production tests; however, the production stimuli were embedded in a sentence frame and participants produced them in a slightly different sentence frame. Production scores reflect how the learners' productions of each category were recognized by the English Model. Only data from the 22 participants who completed both the identification and production tests were used to calculate these means. Furthermore, only the first repetition of each production stimulus item for each learner was used, reflecting the fact that only one repetition of each item was presented in the English vowel identification test. Figures 4.25 and 4.26 illustrate responses to Voice 1 stimuli at each Time; Figures 4.27 and 4.28 illustrate responses to Voice 2 stimuli at each Time.

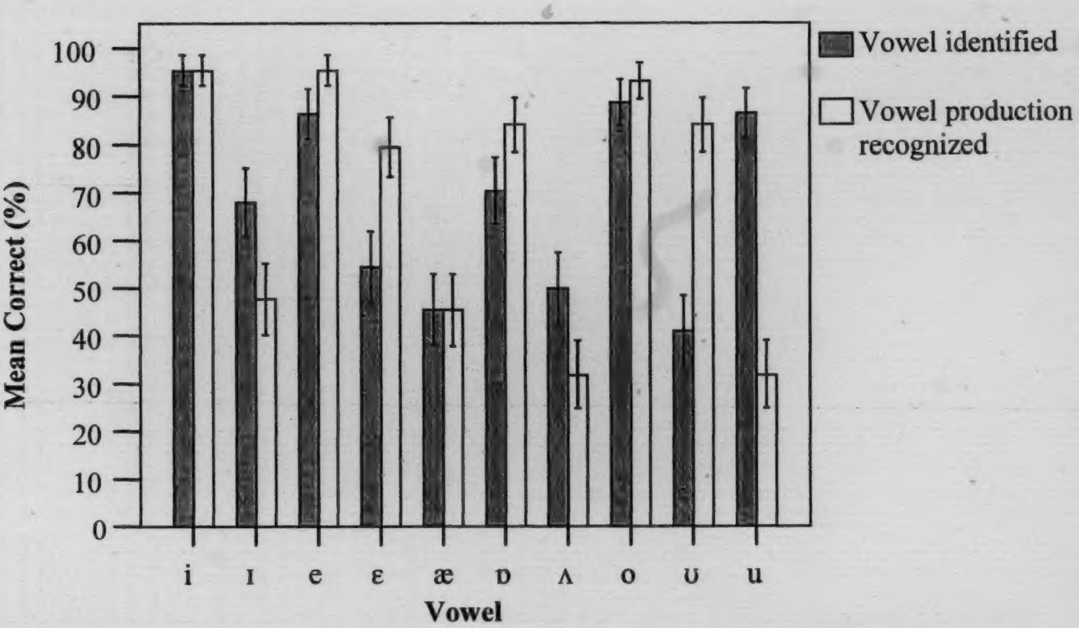


Figure 4.25. Comparison of average vowel perceptual identification and production recognition scores in response to Voice 1 /b, pV/ stimuli at Time 1. Error bars represent standard errors.

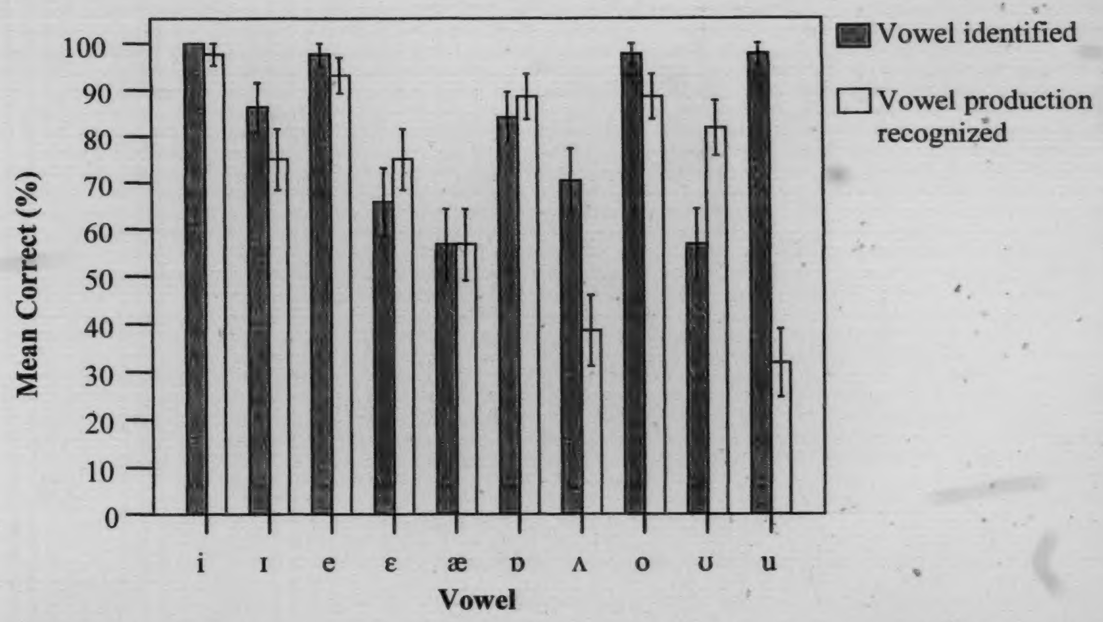


Figure 4.26. Comparison of average vowel perceptual identification and production recognition scores in response to Voice 1 /b, pV/ stimuli at Time 2. Error bars represent standard errors.

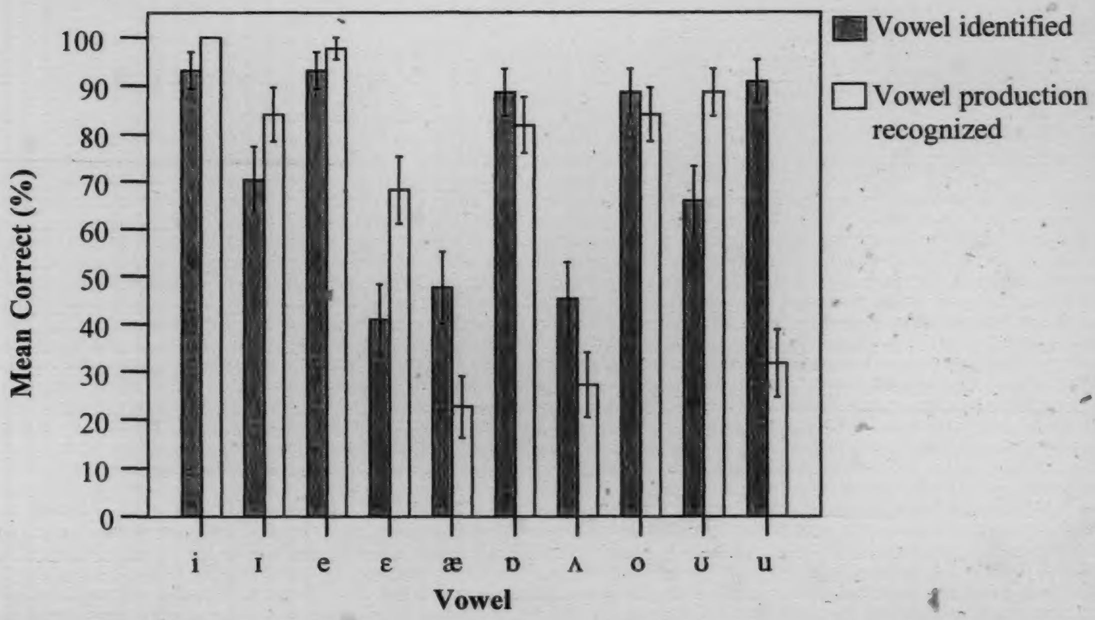


Figure 4.27. Comparison of average vowel perceptual identification and production recognition scores in response to Voice 2 /b, pV/ stimuli at Time 1. Error bars represent standard errors.

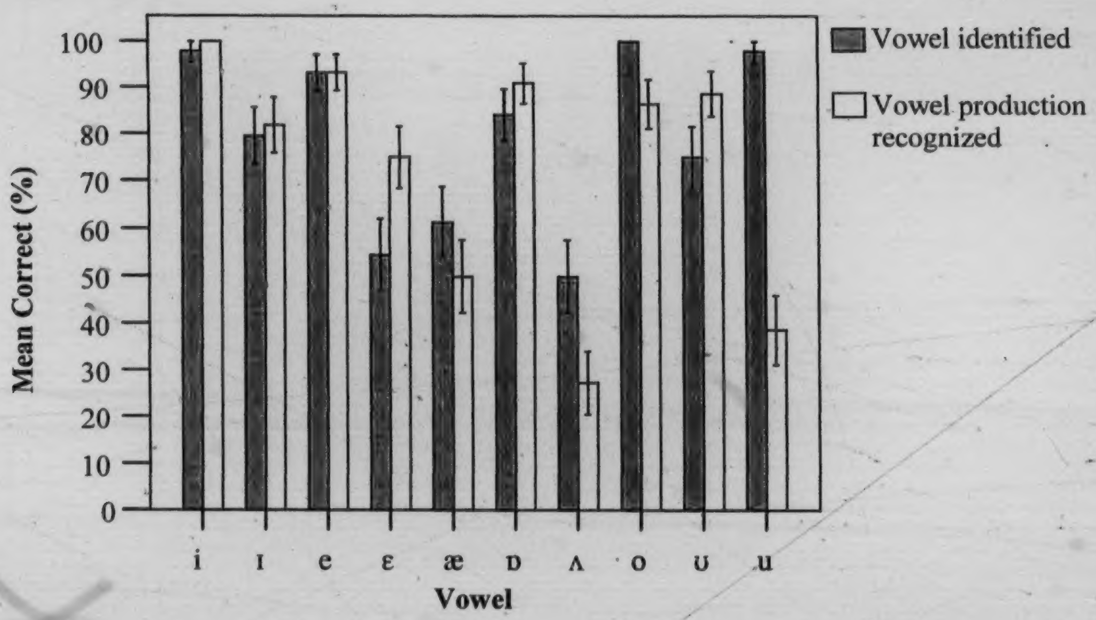


Figure 4.28. Comparison of average vowel perceptual identification and production recognition scores in response to Voice 2 /b, pV/ stimuli at Time 2. Error bars represent standard errors.

Caution should be exercised in interpreting the contrasts between identification and production recognition scores illustrated for Time 1. The Time 1 production test was conducted prior to the beginning of training, while the Time 1 identification test was conducted after four training sessions; recall, however, that during the initial training sessions, although some learning most likely occurred, the participants were still becoming accustomed to the training task.

The contrast between Time 2 identification and production recognition tests provides a more valid comparison than those for Time 1; the Time 2 production test was administered the day after the Time 2 identification test, with no intervening training sessions. Differences in mean accuracy scores on the identification and production tests at Time 2 suggest that for most English vowels, the learners' ability to identify a vowel is similar to their ability to produce that vowel. However, while a relationship between identification and production recognition scores seems to exist for most vowels, it is not the case for all vowels. At Time 2, mean scores for English /ʌ/ and /u/ productions in particular are much worse than the same vowels' mean identification scores, regardless of stimulus voice. A closer examination of the L2 data revealed that errors in the recognition of English /u/ productions were largely in the direction of English /ʊ/; errors in the recognition of English /ʌ/ productions were largely in the direction of English /ɒ/.

Figures 4.29 – 4.32 illustrate English /u/-/ʊ/ recognition patterns in terms of each L2 production token's F1/F2 values at 20% and 70% points of each vowel production's duration. Separate plots are provided for L2 productions in response to each stimulus voice at each time. Frequency values are normalized for pitch based on an alternate method reported by Nearey and Assmann (1986, p 1305, note 5). Those authors found that this alternate method yielded similar results to their main method, which serves as the basis of the pattern recognition models used in this thesis. The explicit normalization step is adopted here because it makes it easier visualize results in 2-dimensional subspaces and reduces variability among speakers.

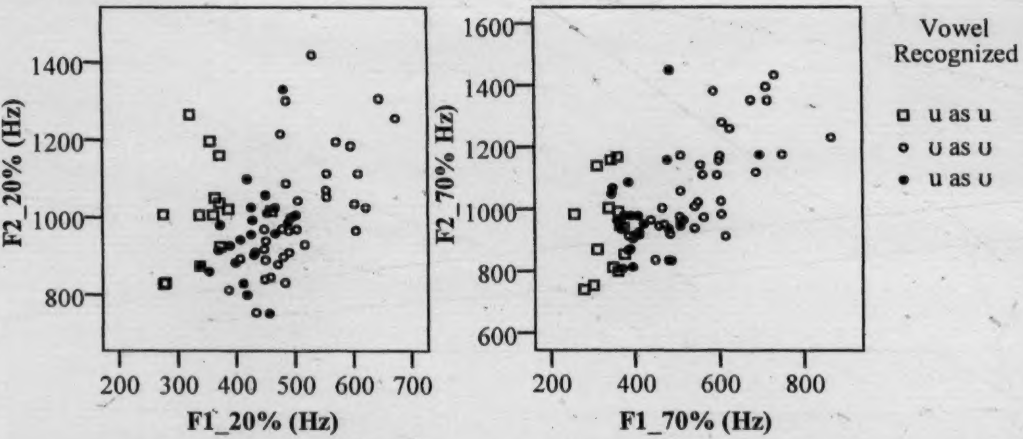


Figure 4.29. F0 normalized production measures for responses to Voice 1 /u/ and /ʊ/ stimuli at Time 1. Scatterplots illustrate F1/F2 values taken from 20% and 70% points of vowel length. Marks indicate whether the vowel was recognized by the English Model as the intended vowel or the competing vowel in the /u/-/ʊ/ contrast.

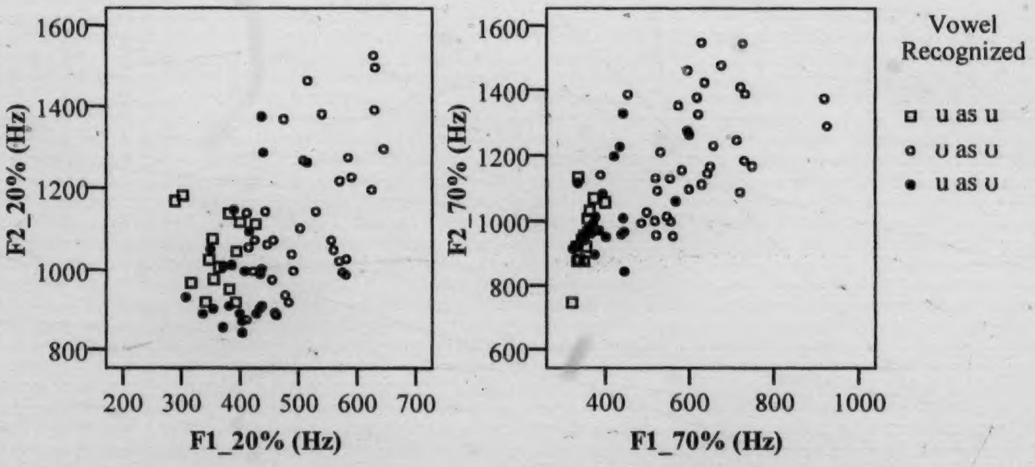


Figure 4.30. F0 normalized production measures for responses to Voice 1 /u/ and /ʊ/ stimuli at Time 2. Scatterplots illustrate F1/F2 values taken from 20% and 70% points of vowel length. Marks indicate whether the vowel was recognized by the English Model as the intended vowel or the competing vowel in the /u/-/ʊ/ contrast.

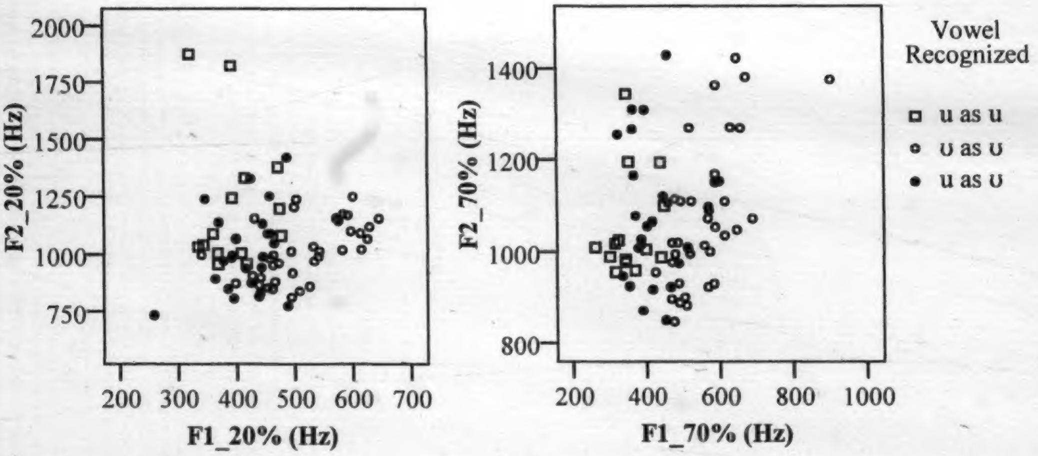


Figure 4.31. F0 normalized production measures for responses to Voice 2 /u/ and /ʊ/ stimuli at Time 1. Scatterplots illustrate F1/F2 values taken from 20% and 70% points of vowel length. Marks indicate whether the vowel was recognized by the English Model as the intended vowel or the competing vowel in the /u/-/ʊ/ contrast.

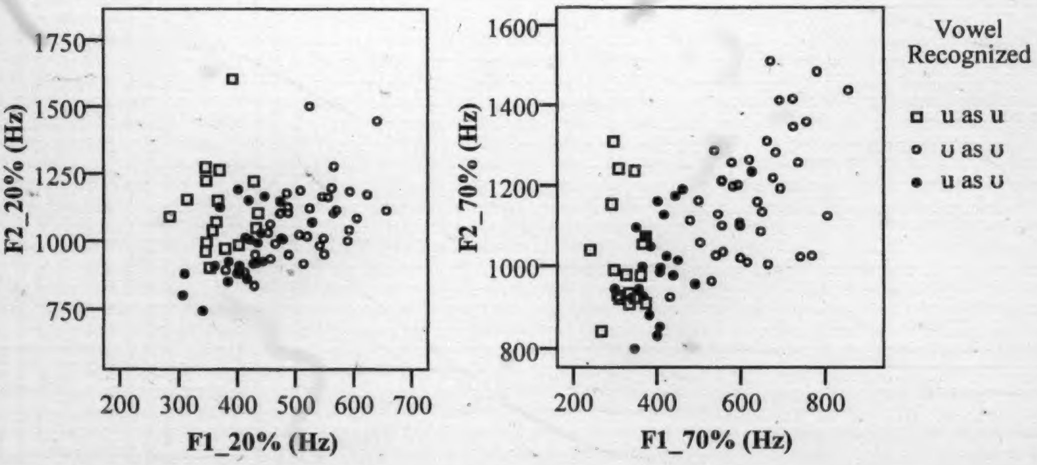


Figure 4.32. F0 normalized production measures for responses to Voice 2 /u/ and /ʊ/ stimuli at Time 2. Scatterplots illustrate F1/F2 values taken from 20% and 70% points of vowel length. Marks indicate whether the vowel was recognized by the English Model as the intended vowel or the competing vowel in the /u/-/ʊ/ contrast.

The F1/F2 scatter plots for learner productions of English /u/ and /ʊ/ indicate similar patterns across time and stimulus voice. When productions of /u/ were recognized by the English Model including duration as /ʊ/, they generally did not share the same spectral properties as learner productions of /ʊ/ that were accurately recognized as /ʊ/. Rather, the learners appear to be making a distinction between /u/ and /ʊ/, but the boundary for some learners appears to be within the English Model's /ʊ/ category; therefore some productions of English /u/ were recognized as /ʊ/. Conversely, no intended productions of /ʊ/ were incorrectly recognized as /u/. Another observation is that the F1/F2 separation between the learners' correctly recognized /u/ productions and those recognized as /ʊ/ is greatest at the beginning of the vowel (the 20% point), but less obvious at the end of the vowel (70% point). This pattern is slightly stronger at Time 2.

Vowel duration differences between /u/ productions that were correctly recognized as /u/, versus those that were incorrectly recognized as /ʊ/, indicate no clear difference between correct and incorrectly recognized productions. That is, whether the English Model recognized the intended /u/ production as /u/ or as /ʊ/, the production's duration was much longer than for intended and correctly recognized productions of /ʊ/. These duration differences are provided in Table 4.8. Interestingly, although still relatively long, the mean duration of /u/ productions which were incorrectly recognized as /ʊ/ is shorter than the mean duration of productions of /u/ that were correctly recognized as /u/, regardless of stimulus voice or time.

Table 4.8. Mean vowel duration and standard deviation for productions of /u/ and /ʊ/, in terms of how the intended vowel was recognized by the English Model. Results are provided separately for responses to each stimulus voice at each time.

		Vowel intended	Vowel recognized	% of intended vowel tokens (n=44)	Mean duration	*Standard deviation*
Response to Voice 1 Stimulus	Time 1	/u/	/u/	31.8	381.1	*122.4*
		/ʊ/	/ʊ/	84.1	234.6	*67.7*
		/u/	/ʊ/	47.7	340.0	*75.7*
		/ʊ/	/u/	0	n/a	*n/a*
	Time 2	/u/	/u/	31.8	532.9	*127.7*
		/ʊ/	/ʊ/	88.6	236.5	*62.5*
		/u/	/ʊ/	50.0	475.5	*87.8*
		/ʊ/	/u/	0	n/a	*n/a*
Response to Voice 2 Stimulus	Time 1	/u/	/u/	31.8	422.0	*118.9*
		/ʊ/	/ʊ/	81.8	238.4	*53.9*
		/u/	/ʊ/	47.7	368.5	*105.0*
		/ʊ/	/u/	0	n/a	*n/a*
	Time 2	/u/	/u/	38.6	569.2	*105.1*
		/ʊ/	/ʊ/	88.6	228.5	*78.7*
		/u/	/ʊ/	52.3	519.1	*95.8*
		/ʊ/	/u/	0	n/a	*n/a*

Figures 4.33 – 4.36, below, illustrate English /ɒ/-/ʌ/ recognition patterns in terms of each token's F1/F2 values at 20% and 70% points of each vowel production's duration. Separate plots are provided for L2 productions in response to each stimulus voice at each time. As previously, frequency values are normalized for pitch based on an alternate method reported by Nearey and Assmann (1986, p 1305, note 5).

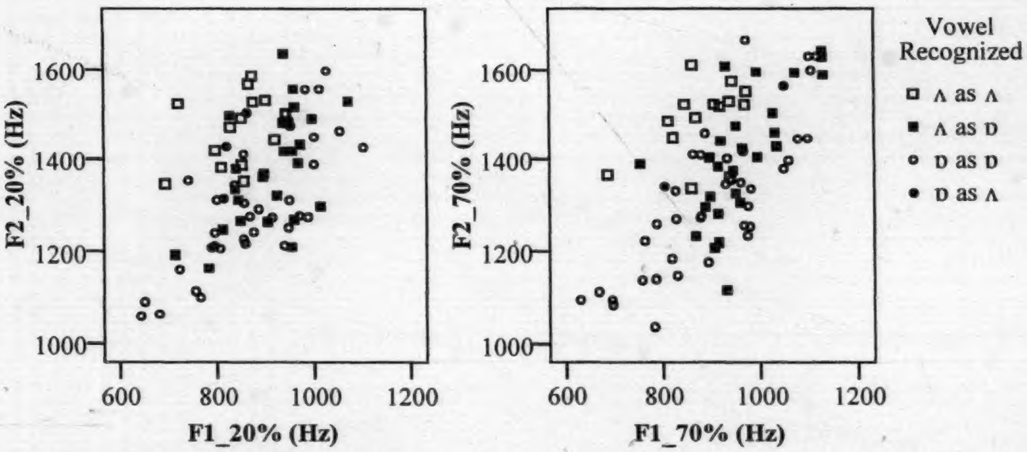


Figure 4.33. F0 normalized production measures for responses to Voice 1 /ɒ/ and /ʌ/ stimuli at Time 1. Scatterplots illustrate F1/F2 values taken from 20% and 70% points of vowel length. Marks indicate whether the vowel was recognized by the English Model as the intended vowel or the competing vowel in the /ɒ/-/ʌ/ contrast.

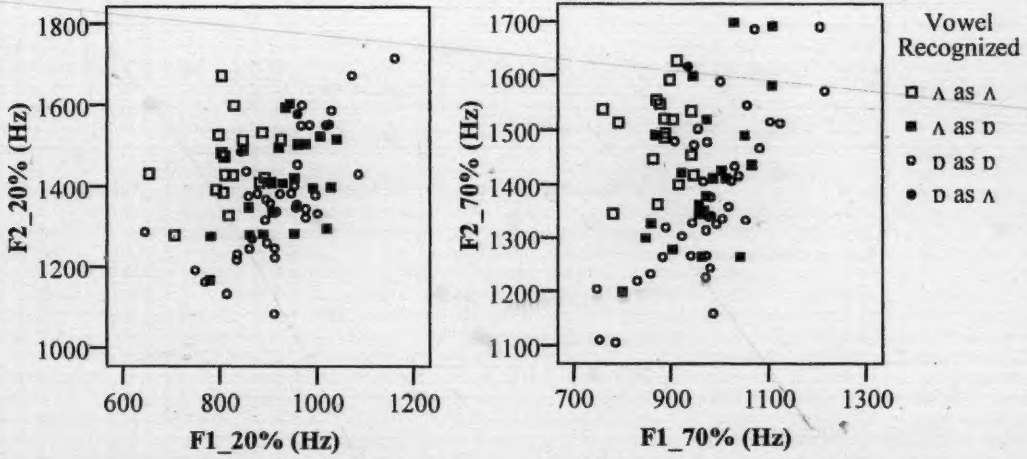


Figure 4.34. F0 normalized production measures for responses to Voice 1 /ɒ/ and /ʌ/ stimuli at Time 2. Scatterplots illustrate F1/F2 values taken from 20% and 70% points of vowel length. Marks indicate whether the vowel was recognized by the English Model as the intended vowel or the competing vowel in the /ɒ/-/ʌ/ contrast.

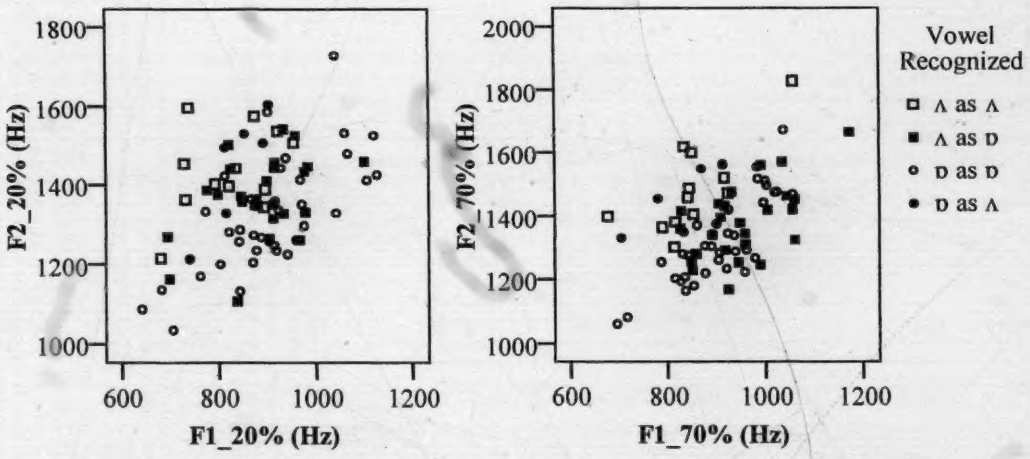


Figure 4.35. F0 normalized production measures for responses to Voice 2 /ɒ/ and /ʌ/ stimuli at Time 1. Scatterplots illustrate F1/F2 values taken from 20% and 70% points of vowel length. Marks indicate whether the vowel was recognized by the English Model as the intended vowel or the competing vowel in the /ɒ/-/ʌ/ contrast.

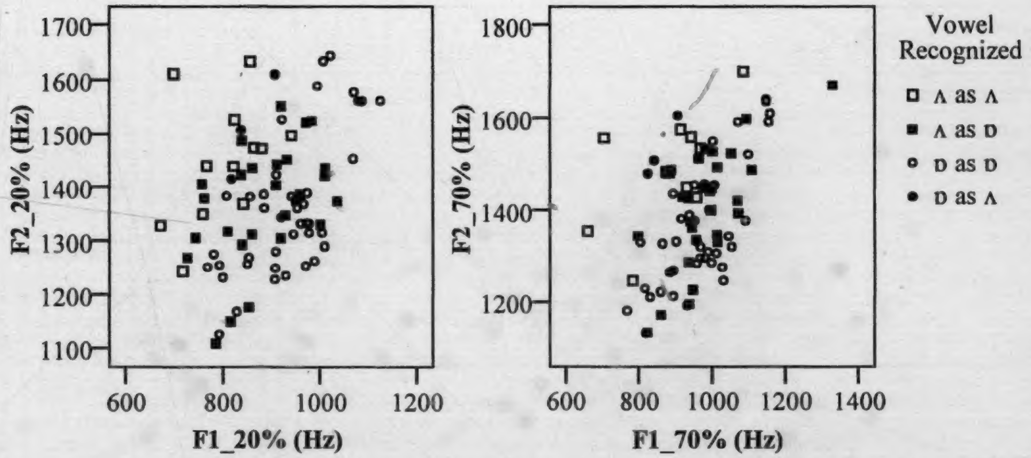


Figure 4.36. F0 normalized production measures for responses to Voice 2 /ɒ/ and /ʌ/ stimuli at Time 2. Scatterplots illustrate F1/F2 values taken from 20% and 70% points of vowel length. Marks indicate whether the vowel was recognized by the English Model as the intended vowel or the competing vowel in the /ɒ/-/ʌ/ contrast.

The F1/F2 scatter plots for learner productions of /ʌ/ and /ɒ/ indicate that only a very limited spectral separation exists between learner productions of /ɒ/ that were correctly recognized as /ɒ/ and learner productions of /ʌ/ that were incorrectly recognized as /ɒ/. That is, many incorrectly recognized learner productions of /ʌ/ were very similar in the F1/F2 dimension to the learners' productions of /ɒ/. Separation between learner productions of /ʌ/ that were recognized as /ɒ/ versus their intended and correct productions of /ɒ/ may be slightly greater at the end of the vowel (70% mark) than at the beginning (20% mark) and slightly greater at Time 2 than at Time 1.

Vowel duration differences between /ʌ/ productions that were correctly recognized as /ʌ/, versus those that were incorrectly recognized as /ɒ/, indicate no clear difference between correct and incorrectly recognized productions; both have relatively short durations. That is, whether the English Model recognized the intended /ʌ/ production as /ʌ/ or as /ɒ/, the duration was much shorter than for intended and correctly recognized productions of /ɒ/. These duration differences are illustrated in Table 4.9. Interestingly, although still relatively short, the mean duration of learner productions of /ʌ/ which were incorrectly recognized as /ɒ/ was longer than the mean duration of productions of /ʌ/ that were correctly recognized as /ʌ/, regardless of stimulus voice and time.

Table 4.9. Mean vowel duration and standard deviation for productions of /ɒ/ and /ʌ/, in terms of how the intended vowel was recognized by the English Model. Results are provided separately for responses to each stimulus voice at each time.

		Vowel intended	Vowel recognized	% of intended vowel tokens (n=44)	Mean duration	*Standard deviation*
Response to Voice 1 Stimulus	Time 1	/ɒ/	/ɒ/	84.1	363.1	*62.3*
		/ʌ/	/ʌ/	31.8	203.3	*60.4*
		/ɒ/	/ʌ/	6.8	325.3	*137.3*
		/ʌ/	/ɒ/	59.1	227.0	*72.1*
	Time 2	/ɒ/	/ɒ/	81.8	458.0	*75.6*
		/ʌ/	/ʌ/	27.3	208.5	*61.1*
		/ɒ/	/ʌ/	13.6	427.7	*80.3*
		/ʌ/	/ɒ/	59.1	213.3	*68.3*
Response to Voice 2 Stimulus	Time 1	/ɒ/	/ɒ/	88.6	423.5	*80.2*
		/ʌ/	/ʌ/	38.6	206.2	*55.4*
		/ɒ/	/ʌ/	2.3	400.0	*n/a*
		/ʌ/	/ɒ/	50.0	242.5	*81.8*
	Time 2	/ɒ/	/ɒ/	90.9	544.8	*86.5*
		/ʌ/	/ʌ/	27.3	188.5	*32.6*
		/ɒ/	/ʌ/	6.8	477.3	*87.3*
		/ʌ/	/ɒ/	61.4	199.5	*62.9*

Differences in the F1/F2 dimension for correctly and incorrectly recognized productions of English /u/ and /ʌ/ indicate that in spectral dimension, separation between competing /u/-/ʊ/ and /ʌ/-/ɒ/ is limited. However, in the duration dimension, separation between the members in each of these contrasts is very large. This suggests that on the identification task, relying on duration might allow learners to correctly identify members of these pairs, while on the production task, duration would be insufficient. Examining the duration of each stimulus item used to elicit these vowels, durational differences are clear (see Table 4.10 below). The nominally long vowels in each stimulus pair are much longer in duration than the nominally short vowels in each pair. Comparing duration differences across stimulus voices, these results indicate that while /ʌ/ and /ʊ/ have similar durations when produced by either Voice 1 or Voice 2, the durations of /u/ and /ɒ/ productions are much longer when produced by Voice 2 than when produced by Voice 1.

Table 4.10. Vowel duration for each stimulus CV, by stimulus voice.

Stimulus CV	/bu/	/pu/	/bʊ/	/pʊ/	/bɒ/	/pɒ/	/bʌ/	/pʌ/
Voice 1	346	238	192	170	320	274	220	186
Voice 2	454	410	214	174	524	388	240	170
Average across consonant onsets and voices	362		187.5		376.5		204	

Comparing this fact with L2 productions of these vowels previously shown in Tables 4.8 and 4.9, a similar pattern emerges; with the exception of /ʌ/, the mean duration of L2 vowel productions for these vowels in response to Voice 2 stimuli are substantially longer than the mean duration of L2 vowel productions in response to Voice 1 stimuli. This suggests that the learners are preserving relative vowel duration in their elicited imitation of these syllables.

These results concerning vowel duration provide a possible explanation for the learners' higher accuracy rates on the identification task relative to the production task. While duration by itself may be sufficient for identifying members of these spectrally similar English vowel pairs, in production duration will not suffice. The speaker must also correctly produce spectral properties associated with each vowel. It should also be recognized that duration was not included as a variable in the pattern recognition model that was used as the primary assessment of the L2 productions. As can be seen in Tables A8.8 and A8.9 in Appendix 8, when vowel duration is included as a variable, the recognition scores for /ʌ/ are higher although still poor. In contrast, including duration as a variable in the production recognition model does not increase the production recognition scores for /u/ at all, suggesting again that learners may be using duration as a cue for both identification and production of /u/ with less attention paid to its spectral properties.

Finally, it should be noted that the difficulty these L2 English learners face in discerning spectral differences between /ɒ/ and /ʌ/ in particular may stem from relative spectral ambiguity that exists within this particular contrast, even for native speakers. When vowel duration was excluded as a variable in the English Model developed in Chapter 3, 25% of /ɒ/ productions were recognized as /ʌ/, while 12.5% of /ʌ/ productions were recognized as /ɒ/. Conversely, when duration was included, error rates were only 5% and 2.5% respectively. Recall that the English model was trained and tested on 20 voices. The stimuli used in the identification and production tests also provide evidence of ambiguity. Even though carefully articulated, the vowel in the Voice 2 /pɒ/ stimulus item was incorrectly recognized as /ʌ/ when tested against the English Model that excluded vowel duration as a variable. Furthermore, although not incorrectly recognized in absolute terms, the a posteriori probabilities of the /ʌ/ test stimuli being classified as /ɒ/ increase substantially when vowel duration is excluded as a variable in the English Model (Compare Tables A3.1 and A3.2 with Tables A3.5 and A3.6 in Appendix 3).

4.4. Discussion

In the introduction to the training study reported in this chapter, I posed five research questions related to the effect of English vowel identification training on L2 perception and production. In the following section, I will answer each research question in turn.

4.4.1. Perception

The first research question asked, when learners are presented with different types of training in phonetic contrasts, what global effects are there on the development of their L2 English vowel identification ability? From the results of this training experiment, it appears that all training conditions led to significant improvement in the learners' ability to identify English vowels over time. Furthermore, there did not appear to be any substantial difference between training conditions. This means no one group's stimuli turned out to be much better or worse than another's when it came to overall learning.

The fact that there was significant improvement after just four short training sessions indicates that this form of training with multiple contrasts and feedback on choices is very successful. The fact that there was not a significant difference in performance across conditions does not necessarily mean that these types of modifications to training stimuli are not potentially meaningful. It may be that the training task was so successful in orienting the learners' attention to meaningful acoustic cues that additional information in the form of stimulus modification had little measurable effect. In other words, given the degree to which learners were focused on the stimuli, the effect of the relative salience of particular tokens was washed out. It could be that in conditions where attention is not so effectively oriented to the training stimuli, differences may be detectable. For example, had the task been designed to draw the learners' attention away from the target vowels, differences in stimuli type may have had a larger impact. An example of such a design was used by Guion and Pederson (2004), who oriented learners' attention to consonants in the stimuli set and then tested their ability to identify vowel contrasts that were also part of the stimuli set. Although a different design might answer questions regarding the effect of different types of modified stimuli, since the goal of instruction is to maximize learning, it appears that in contexts where attention can be dedicated to the vowel learning task, using unmodified English vowels will be just as effective as using modified stimuli.

The second research question asked whether vowel identification training in one CV context will transfer to identification ability in another CV context. The answer is yes. Although there are clear differences in performance depending on what onset consonant is used, all contexts improved over time, despite training that focused on a single context only. There is a weak indication that transfer to /g, kV/ contexts is larger than to /z, sV/ contexts. Perhaps this is due to the degree of similarity between /b, pV/-/g, kV/ versus /b, pV/-/z, sV/. In the former, the manner of articulation is the same; all are stops. In the latter, the manner is different, contrasting stops with fricatives. The effect of frication noise preceding the vowel may cause perceptual difficulties for L2 learners.

The third research question asked to what extent the effect of identification training transfers to new tokens produced by a familiar voice and new tokens produced by a new voice. The answer is that overall transfer to new tokens occurs equally, whether the voice is familiar or unfamiliar. The most likely reason for this finding is that the high degree of variability provided in the training stimuli was sufficient to transfer to new tokens and speakers. This effect of high variability training was previously established by Logan, Lively and Pisoni (1991) as well as Pisoni and Lively (1995). At the same time, the results of the current study also indicate that for particular vowels, performance can vary by stimulus voice. This suggests that differences in productions of the same category by different speakers may interact differently with the L2 learners' L1 categories, or with other new categories emerging in the L2 system.

The fourth research question asked whether evidence of improvement in perception would still be detectable one month after training was completed. The results of the delayed post-tests indicate that the ability to identify English vowels had not significantly declined, yet neither had it continued to improve in the absence of instruction. This suggests that the improvement during training was the result of training itself and not general learning apart from training; although some small improvement may have occurred, measurable improvement in the absence of training could not be detected in such a short period of time. Conversely, there is a possibility that the end of training coincided with a point when general learning began to asymptote.

Apparent decreases in the identification rates of some vowels may have been caused by the participants' decreasing familiarity with the task. At Time 2, they may have had stronger connections between category and identification symbol than later at the delayed post-test, when they had experienced a one month break away from the computer program. In fact several participants did indicate that although I allowed them a few minutes to practice before beginning the test, they had forgotten what categories went with which symbol, particularly for the categories that were not strong to begin with. For example, the sound-symbol relationship for the English vowel /i/ was reinforced from nearly the beginning of the training period and therefore that connection may have been stronger than for categories that were more difficult, and thus received less reinforcement during training.

4.4.2. Production

The fifth research question asked to what degree identification training transfers to production. It appears from this study that identification training on ten English vowel categories does transfer to production. However, while there is strong evidence that this transfer occurs in the /b, pV/ context on which learners were trained, a statistically significant improvement in different CV contexts is not evident. This does not necessarily mean that learning did not occur. There are other explanations for why improvement was not detected. First, recall that on the identification training tests, learners improved in all three contexts, but not to the same extent. Since the mean differences between pre and post training on the identification tests were much larger than they were for the production tests, this may suggest that learning is too limited to yet be detectable in production. Another possibility is that learners had begun practicing the articulation of English /b, pV/ contrasts during training and were therefore more prepared to be tested on them in production. During the course of training, I often observed participants sub-vocalizing (and sometimes even vocalizing) the sounds they were hearing. Since they were not trained on /g, kV/ and /z, sV/ contexts, they did not have opportunity to practice the articulation of the vowels in these contexts. If production lags behind perception, this finding may indicate that although they were better able to identify English vowels in new contexts, they needed practice to produce them more accurately. This suggests that incorporating an overt repetition task into the identification training may facilitate transfer to production. However, research is needed to determine if overt repetition before at least some identification training has occurred may actually distract attention from the perceptual training.

The failure to detect significant improvement in /g, kV/ and /z, sV/ contexts may also be related to the fact that these productions were tested on a statistical model that had only been trained on English /b, pV/ productions. Although this model relied on measurements taken from 20% into each vowel token's total duration, it is still possible that some effect of transition from the preceding consonant was unavoidably incorporated into the model, slightly biasing it against accurate identification of /g, kV/ and /z, sV/ contexts. This latter possibility seems to be the least likely explanation, since Tukey HSD tests indicated that mean correct identification rates for the /b, pV/ context versus the other contexts was not significant at Time 1. This seems to suggest that the statistical model should also be able to detect relative improvement, even if absolute correct identification scores for the /g, kV/ and /z, sV/ tokens were not as accurate as for the /b, pV/ tokens. Furthermore, when the CV test stimuli were tested against the English Model, all but one of the /g, kV/ and /z, sV/ stimuli were recognized correctly.

Finally, the ability to identify particular vowels on the identification test was not necessarily mirrored in the speakers' ability to produce the same vowels. The two most obvious examples were English /u/ and /ʌ/. Learners' identification scores for these vowels were much stronger than their production recognition scores, even when vowel duration was included as a variable in the production recognition model.

On the identification test, participants were able to accurately identify /u/ most of the time. In production, however, /u/ had one of the lowest accuracy rates. Learners were also better able to identify /ʌ/ than to produce a recognizable form of it in production. However, even in identification, /ʌ/ was not particularly strong vis-à-vis other vowel categories.

One possible explanation for why /u/ and /ʌ/ were easier to identify than they were to produce may be that on the identification task, learners could attend to duration as a primary cue to identifying the difference between both /u/-/ʊ/ and /ɒ/-/ʌ/; the learners showed an ability to use duration to distinguish between members of each pair in production. However, while attending to duration would potentially facilitate success on the identification test, if that were the primary cue being used, learners might fail to learn the relevant spectral dimensions associated with each vowel. This then could explain why in production the same vowels were recognized as being produced incorrectly with or without duration used in the recognition model.

Finally, it should be mentioned that given the differences in the nature of the identification and production tests, some differences in performance on each test may be expected. On the identification test, each CV stimulus was presented in isolation. Consequently, learners' attention may have been better oriented to picking up the relevant acoustic information needed for successful vowel identification. In contrast, for the production test each CV stimulus was embedded in the carrier phrase, "The next word is _____." This may have made detection of spectrally relevant information more difficult. However, while this might be a reasonable explanation for vowels that were poorly identified and produced (i.e., /ʌ/), it seems like a less appealing explanation for vowels that were well identified, but poorly produced (i.e., /u/); the other well identified vowels were also produced with high accuracy, suggesting that in at least some instances, any adverse effect of the carrier sentence used to elicit production was minimal.

Chapter 5. Testing predictions of L2 speech learning models

The two most influential models of L2 speech perception, Best's (1995) Perceptual Assimilation Model (PAM) and Flege's (1995) Speech Learning Model (SLM), make specific predictions regarding the type of difficulty L2 learners will face in acquiring L2 phonological categories in terms of those categories' interactions with pre-existing L1 phonological categories. In this chapter, I analyze the identification and production data from the training study described in the preceding chapter in terms of these predictions. However, while I am framing my analysis in terms of PAM and SLM, where I use these labels, as I have stated previously, it should be understood that they are being interpreted in terms of the statistically defined measures of crosslinguistic similarity outlined in Chapters 2 and 3. Hence, when I describe a category in the L2 as being 'similar' to an L1 category, I mean that many of the production tokens that comprise that L2 category are statistically similar to many of the production tokens that comprise the L1 category[16]. When I describe a category in the L2 as being 'new' for a learner, I mean that the most of the production tokens that comprise that category are relatively distant, statistically, from production tokens of any L1 categories[17]. In a later incarnation of PAM, Best, McRoberts and Goodell (2001) move away from reference to phonetic or phonological categories, and instead refer to "functional equivalence classes", maintaining that there is not any ecologically grounded basis for making assumptions regarding cognitive processing or mental representations. I continue to use the term 'category' because it has currency in the literature, and provides a necessary delimitation for the amount of data being analyzed. However, since the Metamodel approach to crosslinguistic similarity introduced in Chapter 2 and 3 applies to individual tokens, it may be viewed as even more loosely comprising categories than does Best et al.'s (2001) notion of "functional equivalence classes". Because the general claims of PAM and the SLM are interpretable in my statistical approach to crosslinguistic similarity, and indeed have influenced its development, I continue to use the terms. However, for the purposes of this study, they should always be understood to have a statistical flavor.

---

[16] Applying this to Mandarin and English vowel categories, they were defined as 'similar' in Chapter 3, when a relatively large proportion of tokens of given Mandarin and English categories were recognized as members of the opposing language category, or at least had some reasonable probability (i.e., > .05 APP) of being recognized as members of the opposing language category by the Metamodel.

[17] Applying this to Mandarin and English vowel categories, English categories were defined as 'new' in Chapter 3, when few if any tokens of an English category were recognized as members of a competing Mandarin category. Additionally, few tokens of the 'new' category had even a reasonable probability (e.g., > .05 APP) of being recognized as a member of a Mandarin category by the Metamodel.

5.1. Predictions

The following predictions were made on the basis of the Metamodel comparison of Mandarin and English vowel categories in Chapter 3, but prior to analyzing the L2 learner data from the training study. Consequently, they represent a strong test of my claims concerning the nature of Mandarin-English vowel similarity as defined by the Metamodel.

5.1.1. Identification confusion patterns

In Chapter 3, I concluded that PAM provides a more useful framework than the SLM for making predictions regarding L2 perceptual confusion patterns in the context of an L2 categorical identification task. In a vowel identification task such as that used in this study, similarity between Mandarin and English categories may actually facilitate rather than hinder categorical identification of some English vowels. If tokens of an L2 category usually assimilate to a single L1 category, the learner will be able to correctly identify the L2 category by bootstrapping on the similar L1 category. In such cases, the learners' ability to correctly identify an L2 category does not depend on their ability to discriminate small phonetic differences between the similar L1 and L2 categories. The SLM is more concerned with learners' ability to discriminate these small phonetic differences between L1 and L2 categories, something that needs to be tested using a discrimination rather than identification task.

In brief review, PAM posits two assimilation patterns that are relevant to the Mandarin and English vowels examined in this study: 1) direct assimilation of an L2 sound to an L1 category, and 2) assimilation of the L2 sound into the learner's system as an uncategorizable speech sound. Recall that direct assimilation can take two forms, each having a different effect. In one form, a single L2 category may assimilate to a single L1 category. In this case, correct identification of the L2 sound should be facilitated. Alternately, two or more L2 categories may assimilate to a single L1 category. In this case, confusion between the L2 categories is more likely to ensue. For the L2 English identification data, I am interested in testing predictions that direct assimilation of an L2 category to an L1 category will facilitate L2 category identification, while assimilation of two or more L2 categories to a single L1 category will hinder identification. In addition, I am interested in determining what role within-L2 category confusability plays.

On the basis of similarity between English and Mandarin vowel inventories that was established in Chapter 3, using the Metamodel pattern recognition approach to compare vowels, Table 5.1 provides a summary of predictions influenced by PAM's basic claims regarding confusion patterns that should be found in the L2 English vowel identification data. For the purpose of this analysis, these predictions are based on the recognition that in cases of direct assimilation, all tokens of a given L2 category are not equally assimilated to the L1 category. In other words, speech categories are not treated as monolithic wholes. Rather, it is recognized that some tokens of a given category may assimilate to an L1 category more readily than other tokens. In addition, the predictions in Table 5.1 are modified from PAM's claims to include recognition of potential within-L2 category confusability. Although not accounted for by PAM, this modification reflects the fact that some English vowel tokens may be confusable with other English vowel categories (e.g., the /ɪ/-/ɛ/ contrast), as was demonstrated by the results of the English and Metamodel analyses in Chapter 3.

Table 5.1. Predictions regarding L2 English vowel identification patterns based on the Metamodel analysis of Mandarin and English vowels in Chapter 3.

English Vowel	Degree of similarity to Mandarin	Analogous PAM process	Predicted identification difficulty
/i/	high	Direct assimilation to Mandarin /i/	Very little
/ɪ/	very low	Uncategorizable speech sound	Within English confusion with /ɛ/
/e/	moderate/high	Direct assimilation to Mandarin /e/	Very little
/ɛ/	very low	Uncategorizable speech sound	Within English confusion between /ɪ/-/æ/
/æ/	low	Three-category assimilation Partial assimilation to Mandarin /a/ English /ɒ/ is likely to assimilate more often than either English /æ/or /ʌ/	Better performance on /ɒ/ than /æ/or /ʌ/. Confusion between English /æ/-/ɒ/-/ʌ/ due to Mandarin three-category assimilation Also within English confusion between /æ/-/ɛ/ and /ʌ/-/ʊ/
/ɒ/	moderate		
/ʌ/	low/moderate		
/o/	high	Direct assimilation to Mandarin /o/	Very little
/ʊ/	high	Direct assimilation to Mandarin /ɤ/	Very little
/u/	very low	Uncategorizable speech sound	Very little Good within English discrimination

5.1.2. Production confusion patterns

In Chapter 3, I concluded that the SLM is a more useful starting point than PAM for making predictions regarding L2 *production* confusion patterns. In production, substituting an L1 category for a similar L2 category is said to stem from the learners' inability to perceive small phonetic differences between those similar L1 and L2 categories. For the L2 English production data from my study, I am interested in testing the SLM prediction that English categories that are most 'similar' to Mandarin categories are more likely to be produced as a typical member of the similar Mandarin category, while those English categories that are less similar to any Mandarin category are more likely to be produced as a member of an L2 English category. In addition, I test the prediction that improvement from Time 1 to Time 2 should be related to the degree of token-wise similarity between L1 and L2 categories, with more 'similar' categories evidencing less improvement than 'new' categories. That is, when a larger number of tokens of an L2 category are similar to an L1 category, the amount of exposure and attention necessary to begin forming a new category is greater – assuming that only a small portion of tokens in the input of the 'similar' category are recognized as realistically belonging to a potentially 'new' category. Finally, correctly identified L2 productions of 'new' English vowels should be closer to the intended L2 English category centre than are productions of more 'similar' English vowels. On the basis of similarity between English and Mandarin vowel inventories established in Chapter 3, Table 5.2 provides a summary of more specific SLM influenced predictions regarding the L2 production data.

As with the predictions regarding identification, these predictions incorporate degrees of crosslinguistic similarity, rather than relying on a binary 'similar' vs. 'new' distinction. In addition, these predictions are modified from basic SLM predictions to include recognition of potential within-L2 cross-category confusions – something that is not explicitly accounted for by the SLM. This addition reflects the fact that some English vowels may be confusable with other English vowels (e.g., the /ɪ/-/ɛ/ contrast), as was demonstrated by the results of the English and Metamodel analyses, as described in Chapter 3.

Table 5.2. Predictions regarding L2 English vowel production patterns based on the Metamodel analysis of Mandarin and English vowels in Chapter 3.

English Vowel	'Similar' or 'New'	Predicted production of intended English vowel
/i/	Highly similar	Frequently produced as Mandarin /i/
/ɪ/	New	Produced as English /ɪ/ or another close English category (i.e., /ɛ/ or /æ/)
/e/	Moderately/highly similar	Frequently produced as Mandarin /e/ in production.
/ɛ/	New	Produced as English /ɛ/ or another close English category (i.e., /ɪ/ or /æ/)
/æ/	New/slightly similar	Usually produced as English /æ/ or another close English category (i.e., /ɛ/ or /ʌ/), but sometimes as Mandarin /a/
/ɒ/	Moderately similar	Often produced as Mandarin /a/ but sometimes as English /ʌ/
/ʌ/	Slightly/moderately similar	Sometimes produced as Mandarin /a/, but sometimes as English /ɒ/ or /ʊ/
/o/	Highly similar	Frequently produced as Mandarin /o/
/ʊ/	Highly similar	Frequently produced as Mandarin /ɤ/
/u/	New	Produced as English /u/ or another close English category (i.e., /o/ or /ʊ/)

5.2. Method

5.2.1. L2 English identification data

English vowel identification data from the Natural Vowel, Lengthened Vowel and Generalization tests at Time 1 and 2 of the training study outlined in the preceding chapter were examined to identify the specific types of confusion patterns that are evident for English vowels in /b, pV/ contexts. These confusion patterns are used to evaluate the predictions related to PAM.

5.2.2. L2 English production data

Raw L2 production data outlined in the preceding chapter were reanalyzed using the Mandarin/English Metamodel developed in Chapter 3. Recall that the Metamodel was trained on Mandarin L1 and English L1 vowel productions. For this study, the L2 production data were tested against the Metamodel to identify each production token in terms of its closest English or Mandarin category. The results reflect the number of L2 production tokens that were classified as the intended English vowel as opposed to other English vowels or Mandarin vowels. In addition, a comparison was made of Mahalanobis Distance scores for those tokens of L1 and L2 English that were correctly identified by the English Model.[18] These scores reflect the absolute distance between a production token and each category's centre. The results from the Mahalobonis Distance scores of the L2 production data are used to further evaluate the modified SLM predictions made using the Metamodel APP approach (see Table 5.2.).

5.3. Results

5.3.1. Identification confusion patterns

Natural and Lengthened Vowel tests

Listener identification confusion matrixes for /b, pV/ contrasts on the Natural and Lengthened vowel tests are provided first (Tables 5.3 and 5.4 respectively). Because they are based on the same set of data that was used for creating the Metamodel, these confusions reflect mean responses to stimuli produced by 20 native English speakers for each vowel category. Since the Lengthened Vowel and Natural Vowel tests were not administered to all participants at Time 1, the summary of results for Time 1 only represent those L2 English learners who took each test at Time 1. Both the Natural Vowel and the Lengthened Vowel tests were administered to the entire group at Time 2[19].

---

[18] The English Model trained and tested without vowel duration as a variable was used.

Table 5.3. Listener identification of English /b, pV/ stimuli on the Natural Vowel Test at Time 1 and Time 2. Numbers represent percentage of tokens identified as the category indicated.

						Vowel identified by L2 learner					
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Time 1 (n=15)	Vowel presented as stimuli										
	/i/	**97.2**	1.2	0.2	--	0.8	0.2	0.2	0.2	0.2	--
	/ɪ/	4.2	**57.3**	10.7	18.0	1.0	0.2	2.5	0.5	5.7	--
	/e/	3.5	3.7	**83.7**	2.3	4.5	0.5	0.2	0.3	1.2	0.2
	/ɛ/	--	20.7	5.5	**40.5**	13.3	2.3	7.0	0.8	9.3	0.5
	/æ/	0.2	1.8	3.3	10.8	**32.3**	38.3	12.2	--	1.0	--
	/ɒ/	0.8	0.5	1.3	1.5	9.7	**69.0**	15.0	0.8	1.3	--
	/ʌ/	--	2.8	1.3	2.8	6.0	23.2	**46.0**	1.0	16.3	0.5
	/o/	0.2	1.8	0.3	0.7	0.8	0.2	1.2	**87.0**	4.3	3.5
	/ʊ/	0.5	0.8	0.3	1.7	0.3	0.7	7.5	6.3	**67.3**	14.5
	/u/	0.5	0.5	0.2	0.3	0.7	0.3	0.3	5.5	1.2	**90.5**
Total correct						67.1%					
Time 2 (n=26)	Vowel presented as stimuli										
	/i/	**96.9**	2.3	0.1	0.1	0.1	0.1	0.1	--	--	0.3
	/ɪ/	2.3	**66.8**	4.0	20.6	3.2	0.1	2.2	--	0.7	0.1
	/e/	0.5	1.3	**92.6**	3.6	1.3	0.2	0.2	0.3	--	--
	/ɛ/	0.1	16.7	1.7	**56.6**	14.1	0.9	6.6	--	3.1	0.1
	/æ/	0.1	1.0	0.4	11.3	**48.2**	27.0	11.3	0.3	0.5	--
	/ɒ/	--	0.3	0.3	0.1	7.1	**76.5**	14.6	0.4	0.7	--
	/ʌ/	0.2	0.8	--	2.5	5.7	19.4	**53.6**	0.2	17.5	0.2
	/o/	--	0.1	0.2	0.2	--	--	--	**96.3**	2.4	0.8
	/ʊ/	0.3	0.5	0.1	0.6	1.4	0.2	5.7	3.3	**79.9**	8.1
	/u/	0.6	0.1	0.1	0.2	0.1	0.2	0.2	3.4	3.7	**91.5**
Total correct						75.9%					

---

[19] Tables A7.1 and A7.2 in Appendix 7 provide a contrast between Time 1 and Time 2 for the subsets only. (i.e., those 15 learners who took the Natural Vowel test at both Time 1 and 2 and those 11 learners who took the Lengthened Vowel test at Time 1 and Time 2).

Table 5.4. Listener identification of English /b, pV/ stimuli on the Lengthened Vowel test at Time 1 and Time 2. Numbers represent percentage of tokens identified as the category indicated.

		Vowel identified by L2 learner									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Time 1 (n=11) Vowel presented as stimuli	/i/	**92.7**	5.5	--	--	--	0.9	--	0.2	--	0.7
	/ɪ/	1.8	**57.3**	4.8	20.0	5.7	0.2	6.4	0.2	1.4	2.3
	/e/	0.7	3.0	**84.3**	5.0	2.0	--	0.7	0.5	--	3.9
	/ɛ/	--	27.3	7.7	**42.5**	11.8	0.9	5.5	0.7	3.4	0.2
	/æ/	--	3.0	4.3	8.9	**38.9**	29.3	14.3	0.7	0.5	0.2
	/ɒ/	--	0.2	--	0.5	6.8	**79.3**	12.5	0.5	--	0.2
	/ʌ/	--	2.3	1.1	5.0	8.4	29.8	**35.0**	0.5	17.7	0.2
	/o/	--	0.2	--	0.2	--	1.4	0.9	**92.3**	4.5	0.5
	/ʊ/	--0.2	2.5	0.7	2.0	2.5	0.2	5.7	1.8	**75.9**	8.4
	/u/	0.5	0.2	--	0.7	0.5	2.5	1.1	2.7	5.2	**86.6**
Total correct						68.5%					
Time 2 (n=26) Vowel presented as stimuli	/i/	**98.4**	1.4	0.1	--	--	--	0.1	--	--	--
	/ɪ/	1.6	**62.2**	10.5	17.4	5.6	0.1	1.3	--	1.2	0.2
	/e/	0.1	1.3	**94.8**	1.8	1.1	0.1	0.4	--	0.4	0.1
	/ɛ/		15.8	3.5	**53.0**	20.4	0.8	4.3	0.4	1.9	--
	/æ/	0.3	1.4	0.5	8.3	**57.1**	21.2	10.3	--	1.0	--
	/ɒ/	0.1	0.1	--	0.3	7.6	**81.0**	9.7	0.6	0.7	--
	/ʌ/	0.1	0.7	0.2	2.8	5.6	27.6	**46.1**	0.2	16.8	--
	/o/	0.2	--	0.1	--	0.4	0.1	0.3	**96.8**	1.7	0.4
	/ʊ/	0.1	0.6	--	1.2	1.1	0.3	7.3	2.3	**79.9**	7.3
	/u/	0.9	0.1	--	--	0.3	0.2	0.2	2.9	2.9	**92.6**
Total correct						76.2%					

Given the degree of variation across such a large number of voices and items in the Natural and Lengthened vowel tests, potential differences in listener responses to particular voices are numerous and are therefore not individually highlighted.

The results illustrated in Tables 5.3 and 5.4 indicate that overall confusion rates on the Natural Vowel and Lengthened Vowel tests improved between Times 1 and 2.

However, despite improvement, the confusion patterns in terms of each vowel category's relative degree of confusability with other categories remained largely the same. Time 2 results, which include all 26 participants, indicate that the vowels /i/, /e/, /o/ and /u/ had the highest identification rates, followed closely by /ʊ/ and /ɒ/, which showed only moderate levels of confusion. The vowels /ɪ/, /ɛ/, /ʌ/ and /æ/ demonstrated the highest degrees of confusion with other vowels.

The three English vowels that were predicted to assimilate to the same Mandarin /a/ category (i.e., English /æ/, /ɒ/ and /ʌ/) were confused with each other, as expected. Within this three-way contrast, identification of the English vowel /ɒ/ was most accurate. This was predicted based on the fact that /ɒ/ is far more similar to Mandarin /a/ than are English /æ/ and /ʌ/. The English vowels /ɪ/ and /ɛ/ were also sometimes confused with each other. This was predicted on the basis of their within-English ambiguity that was indicated by both the English Model and the Metamodel. The error patterns found in identification responses to the Natural Vowel and Lengthened Vowel tests support all of the PAM-based predictions, with one minor exception. English /ʊ/ was not identified as accurately as expected, particularly at Time 1. Instead, it showed some confusion with English /ʌ/ or /u/, which was not predicted. However, this English vowel still had very high identification rates relative to the less similar English vowels.

An interesting difference across test types (i.e., Natural versus Lengthened Vowel tests) is evident in confusion patterns for /i/-/ɪ/, /æ/-/ɛ/, /ɒ/-/ʌ/ contrasts. On both tests, the first vowel in each pair was identified far more accurately than the second vowel. However, on the Lengthened Vowel test, identification of the first vowels in these pairs was more accurate than it was on the Natural Vowel test. Conversely, on the Natural Vowel test, identification of the second vowel in each pair was more accurate that it was on the Lengthened Vowel test. Specifically, at Time 2, tokens of /i/, /æ/ and /ɒ/ had identification scores of 98.4%, 57.1% and 81% on the Lengthened vowel test, but only

96.9%, 48.2% and 76.5% on the Natural Vowel Test; at Time 2, tokens of /ɪ/, /ɛ/ and/ʌ/ had identification scores of 66.8%, 56.6% and 53.6%, respectively, on the Natural Vowel test, but only 62.2%, 53% and 46.1% on the Lengthened Vowel test. This may indicate an effect of absolute vowel duration in some learners' vowel identification decisions.

Generalization test

Identification confusion matrixes for /b, pV/ contrasts on the Generalization Test are provided in Tables 5.5 and 5.6, for responses to Voice 1 and Voice 2 stimuli respectively. Examining confusion patterns for each voice separately provides insight into the results reported in the previous chapter, which indicated that there was a significant Voice x Vowel interaction in response patterns. Additionally, pooling the means for responses to two voices would be less informative than is the case when identification is pooled across responses to 20 voices, as was done in presenting results of the Natural and Lengthened Vowel tests. In the context of 20 voices, differences among voices are more likely to be averaged out across stimuli. Furthermore, teasing apart responses to particular voices was not practical for the purposes of this study, given the larger number of voices.

One obvious difference between the Generalization test and the Natural and Lengthened Vowel test results discussed earlier is that overall mean correct responses to both voices on the Generalization test are higher than overall mean correct responses to the Natural and Lengthened Vowel test stimuli. However, as with the Natural and Lengthened Vowel tests, identification confusion patterns in response to the Generalization test stimuli are highly supportive of PAM-based predictions. The vowels /i/, /e/, /o/ and /u/ had the highest identification rates. English /ɒ/ was moderately confusable with other vowels, but less so than it was on the Natural and Lengthened Vowel tests. As with those tests, /ɒ/ was less often confused with other English categories than were /æ/ and /ʌ/, the other English vowels predicted to assimilate to Mandarin /a/. Interestingly, while /æ/ was often confused with /ɒ/ in response to Voice 1 stimuli, this was not the case in response to Voice 2 stimuli.

Table 5.5. Listener identification of English /b, pV/ stimuli on Generalization Test Time 1 and Time 2, Voice 1. Numbers represent percentage of tokens identified as the category indicated.

		Vowel identified by L2 learners									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Time 1 (n=26) Vowel presented as stimuli	/i/	**96.2**	3.8	--	--	--	--	--	--	--	--
	/ɪ/	5.8	**73.1**	3.8	11.5	1.9	--	1.9	--	1.9	--
	/e/	3.8	--	**88.5**	--	5.8	--	--	--	--	1.9
	/ɛ/	--	21.2	3.8	**51.9**	13.5	--	9.6	--	--	--
	/æ/	--	1.9	3.8	11.5	**50.0**	19.2	13.5	--	--	--
	/ɒ/	--	3.8	1.9	3.8	5.8	**73.1**	9.6	--	--	1.9
	/ʌ/	--	7.7	1.9	11.5	1.9	25.0	**51.9**	--	--	--
	/o/	--	--	--	--	--	1.9	--	**90.4**	1.9	5.8
	/ʊ/	--	1.9	--	--	3.8	--	5.8	5.8	**48.1**	34.6
	/u/	3.8	--	--	--	--	--	--	5.8	1.9	**88.5**
Total correct							71.2%				
Time 2 (n=26) Vowel presented as stimuli	/i/	**100**	--	--	--	--	--	--	--	--	--
	/ɪ/	3.8	**88.5**	3.8	3.8	--	--	--	--	--	--
	/e/	1.9	1.9	**96.2**		--	--	--	--	--	--
	/ɛ/	--	13.5	3.8	**63.5**	13.5	--	5.8	--	--	--
	/æ/	1.9	--	--	17.3	**57.7**	11.5	11.5	--	--	--
	/ɒ/	--	--	--	--	3.8	**84.6**	9.6	1.9	--	--
	/ʌ/	--	1.9	--	--	1.9	17.3	**69.2**	--	9.6	--
	/o/	--	--	--	--	--	--	--	**94.2**	3.8	1.9
	/ʊ/	--	--	--	--	--	--	1.9	3.8	**61.5**	32.7
	/u/	--	--	--	--	--	--	--	--	1.9	**98.1**
Total correct							81.3%				

Table 5.6. Listener identification of English /b, pV/ stimuli on Generalization Test Time 1 and Time 2, Voice 2. Numbers represent percentage of tokens identified as the category indicated.

			Vowel identified by L2 learners								
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Time 1 (n=26)	Vowel presented as stimuli /i/	**94.2**	1.9	1.9	--	--	1.9	--	--	--	--
	/ɪ/	5.8	**71.2**	1.9	9.6	--	--	7.7	--	1.9	1.9
	/e/	3.8	--	**92.3**	1.9	1.9	--	--	--	--	--
	/ɛ/	--	28.8	5.8	**40.4**	5.8	1.9	5.8	--	9.6	1.9
	/æ/	3.8	9.6	11.5	5.8	**51.9**	3.8	9.6	1.9	1.9	--
	/ɒ/	--	1.9	--	1.9	9.6	**84.6**	1.9	--	--	--
	/ʌ/	--	1.9	1.9	--	5.8	42.3	**46.2**	--	1.9	--
	/o/	1.9	--	--	--	--	--	1.9	**88.5**	3.8	3.8
	/ʊ/	--	--	--	--	3.8	1.9	1.9	5.8	**71.2**	15.4
	/u/	--	--	--	--	1.9	--	--	3.8	1.9	**92.3**
Total correct						73.3%					
Time 2 (n=26)	Vowel presented as stimuli /i/	**98.1**	--	1.9	--	--	--	--	--	--	--
	/ɪ/	7.7	**80.8**	--	9.6	--	--	1.9	--	--	--
	/e/	--	1.9	**94.2**	--	--	--	1.9	--	--	1.9
	/ɛ/	--	26.9	1.9	**51.9**	13.5	--	5.8	--	--	--
	/æ/	--	1.9	3.8	21.2	**63.5**	5.8	3.8	--	--	--
	/ɒ/	--	--	--	3.8	5.8	**86.5**	3.8	--	--	--
	/ʌ/	--	--	--	1.9	5.8	36.5	**53.8**	--	1.9	--
	/o/	--	--	--	--	--	--	--	**100**	--	--
	/ʊ/	--	--	1.9	--	--	--	1.9	--	**76.9**	19.2
	/u/	--	--	--	--	--	--	--	1.9	--	**98.1**
Total correct						80.4%					

As predicted, the vowels /ɪ/ and /ɛ/ were often confused with each other on the Generalization test, but again, to a lesser extent than on the Natural and Lengthened vowel tests. Interestingly, /ɛ/ was far more likely to be confused with /ɪ/ in response to Voice 2 stimuli than it was in response to Voice 1 stimuli.

As before, the only minor contradiction of predictions in the response data for the Generalization test was the vowel /ʊ/. In response to both voices, relatively high levels of confusion with English /u/ are evident, although confusion in response to Voice 1 was much larger than in response to Voice 2. To a lesser extent, there was also confusion between /ʊ/ and /o/, and again, this confusion was greater for responses to Voice 1.

5.3.2. Production confusion patterns

Metamodel Evaluation

The confusion patterns found for the production data meet most but not all predictions. Production results are provided in Tables 5.7 and 5.8 for L2 productions in response to each Voice at each Time. These results reflect each target L2 English vowel production as it was identified by the Metamodel.

At Time 1, productions of the English vowels /i/, /e/ and /o/, which were deemed most similar to Mandarin categories, were overwhelmingly recognized by the Metamodel as the corresponding Mandarin category. The proportions in which these L2 English productions were recognized as the Mandarin categories vis-à-vis the similar English category were close to the proportions in which L1 Mandarin productions of the same vowels were classified as English vis-à-vis Mandarin categories by the Metamodel (refer to Table 3.5 in Chapter 3). For example, L2 English productions of /i/ at Time 1 were recognized by the Metamodel as English /i/ between 33% and 40% of the time and as Mandarin /i/ between 60% and 65% of the time, depending on the voice used as an elicitation prompt; similarly, Mandarin L1 /i/ productions in the Metamodel analysis in Chapter 3 were identified as English /i/ 27.5% of the time and as Mandarin /i/ 72.5% of the time.

The remaining English vowel that was deemed most similar to a Mandarin category, English /ʊ/, was often recognized in production as Mandarin /ɤ/ as was predicted. However, productions were also frequently recognized as Mandarin /uə/, which was not predicted.

Table 5.7. L2 production data tested on Metamodel excluding vowel duration as variable in response to Voice 1 stimuli. Numbers represent percentage of tokens identified as the category indicated.

		Vowel recognized by Metamodel																
		English										Mandarin						
		/i/$_e$	/ɪ/$_e$	/e/$_e$	/ɛ/$_e$	/æ/$_e$	/ɒ/$_e$	/ʌ/$_e$	/o/$_e$	/ʊ/$_e$	/u/$_e$	/i/$_m$	/e/$_m$	/a/$_m$	/uə/$_m$	/o/$_m$	/ɤ/$_m$	/u/$_m$
Time 1 Target English Vowel	/i/$_e$	33.0	--	2.3	--	--	--	--	--	--	--	64.8	--	--	--	--	--	--
	/ɪ/$_e$	1.1	54.5	9.1	30.7			1.1	--	--	--	--	3.4	--	--	--	--	--
	/e/$_e$	3.4	1.1	21.6	1.1	--	--	--	--	--	--	--	72.7	--	--	--	--	--
	/ɛ/$_e$	--	6.8	--	77.3	10.2	1.1	2.3	--	--	--	--	--	2.3	--	--	--	--
	/æ/$_e$	--	1.1	--	30.7	31.8	4.5	14.8	--	--	--	--	--	17.0	--	--	--	--
	/ɒ/$_e$	--	--	--	--	2.3	19.3	6.8	--	--	--	--	--	68.2	1.1	1.1	1.1	--
	/ʌ/$_e$	--	--	--	1.1	1.1	9.1	12.5	1.1	1.1	--	--	--	73.9	--	--	--	--
	/o/$_e$	--	--	--	--	--	--	--	19.3	--	--	--	--	--	4.5	70.5	2.3	3.4
	/ʊ/$_e$	--	--	--	--	--	2.3	2.3	1.1	15.9	--	--	--	1.1	22.7	9.1	43.2	2.3
	/u/$_e$	--	--	1.1	--	--	--	--	--	3.4	18.2	--	--	--	11.4	5.7	19.3	40.9
Total Correct										30.3%								
Time 2 Target English Vowel	/i/$_e$	40.9	--	1.1	--	--	--	--	--	--	--	58.0	--	--	--	--	--	--
	/ɪ/$_e$	--	69.3	5.7	22.7	1.1	--	--	--	--	--		1.1	--	--	--	--	--
	/e/$_e$	2.3	--	21.6	--	--	--	--	--	--	--	3.4	72.7	--	--	--	--	--
	/ɛ/$_e$	--	9.1	1.1	75.0	10.2	--	2.3	--	1.1	--	--	--	1.1	--	--	--	--
	/æ/$_e$	--	--	--	18.2	55.7	1.1	8.0	--	1.1	--	--	--	15.9	--	--	--	--
	/ɒ/$_e$	--	--	--	--	2.3	28.4	3.4	--	2.3	--	--	--	63.6	--	--	--	--
	/ʌ/$_e$	--	--	--	--	1.1	8.0	22.7	--	3.4	--	--	--	64.8	--	--	--	--
	/o/$_e$	--	--	--	--	--	--	--	19.3	1.1	2.3	--	--	--	3.4	67.0	2.3	4.5
	/ʊ/$_e$	--	--	--	--	--	2.3	13.6	--	19.3	--	--	--	2.3	18.2	--	44.3	--
	/u/$_e$	--	--	--	--	--	--	--	--	2.3	14.8	--	--	--	3.4	1.1	22.7	55.7
Total Correct										36.7%								

Table 5.8. L2 production data tested on Metamodel excluding vowel duration as variable in response to Voice 2 stimuli. Numbers represent percentage of tokens identified as the category indicated.

		Vowel recognized by Metamodel																
		English										Mandarin						
		/i/$_e$	/ɪ/$_e$	/e/$_e$	/ɛ/$_e$	/æ/$_e$	/ɒ/$_e$	/ʌ/$_e$	/o/$_e$	/ʊ/$_e$	/u/$_e$	/i/$_m$	/e/$_m$	/a/$_m$	/uə/$_m$	/o/$_m$	/ɤ/$_m$	/u/$_m$
Time 1 Target English Vowel	/i/$_e$	39.8	--	--	--	--	--	--	--	--	--	60.2	--	--	--	--	--	--
	/ɪ/$_e$	1.1	83.0	2.3	12.5	1.1	--	--	--	--	--	--	--	--	--	--	--	--
	/e/$_e$	--	1.1	33.0	--	--	--	--	--	--	--	--	65.9	--	--	--	--	--
	/ɛ/$_e$	--	15.9	1.1	70.5	3.4	--	6.8	--	--	--	--	1.1	1.1	--	--	--	--
	/æ/$_e$	--	2.3	--	56.8	22.7	1.1	14.8	--	--	--	--	1.1	1.1	--	--	--	--
	/ɒ/$_e$	--	--	--	--	--	20.5	6.8	2.3	--	--	--	--	69.3	--	--	1.1	--
	/ʌ/$_e$	--	--	--	--	3.4	10.2	15.9	--	--	--	--	--	68.2	--	--	2.3	--
	/o/$_e$	--	--	--	--	--	--	--	18.2	--	1.1	--	--	--	5.7	62.5	3.4	9.1
	/ʊ/$_e$	--	--	--	--	--	1.1	--	--	11.4	1.1	--	--	1.1	29.5	3.4	50.0	2.3
	/u/$_e$	--	--	--	--	--	--	--	3.4	5.7	25.0	--	--	--	6.8	2.3	28.4	28.4
Total Correct										34.0%								
Time 2 Target English Vowel	/i/$_e$	46.6	--	--	--	--	--	--	--	--	--	53.4	--	--	--	--	--	--
	/ɪ/$_e$	1.1	84.1	1.1	13.6	--	--	--	--	--	--	--	--	--	--	--	--	--
	/e/$_e$	3.4	--	22.7	--	--	--	--	--	--	--	3.4	70.5	--	--	--	--	--
	/ɛ/$_e$	--	13.6	--	76.1	8.0	--	2.3	--	--	--	--	--	--	--	--	--	--
	/æ/$_e$	--	1.1	1.1	46.6	46.6	1.1	3.4	--	--	--	--	--	--	--	--	--	--
	/ɒ/$_e$	--	[illegible]	--	--	--	30.7	2.3	--	--	1.1	--	--	65.9	--	--	--	--
	/ʌ/$_e$	--	--	--	--	1.1	11.4	13.6	--	2.3	--	--	--	70.5	--	--	1.1	--
	/o/$_e$	--	--	--	--	--	--	--	15.9	3.4	--	--	--	--	2.3	72.7	2.3	3.4
	/ʊ/$_e$	--	--	--	--	--	2.3	4.5	--	18.2	--	--	--	1.1	19.3	1.1	53.4	--
	/u/$_e$	--	1.1	--	--	--	--	--	--	1.1	28.4	--	--	--	9.1	2.3	19.3	38.6
Total Correct										38.3%								

Productions of the English vowel /æ/, which was deemed relatively 'new', though slightly similar to Mandarin /a/, were also recognized as predicted, most often as English /ɛ/ or /ʌ/, but sometimes as Mandarin /a/. Interestingly, there was a large difference in the extent to which intended English /æ/productions were recognized as Mandarin /a/ depending on which stimulus voice was used for production elicitation. When Voice 1 was used, 17% of /æ/ productions were classified as Mandarin /a/ at Time 1 and 15.9% at Time 2. In contrast, when Voice 2 was used, confusion was almost exclusively with other English categories, rather than with the Mandarin category. Only a single token of English /æ/ produced in response to Voice 2 (at Time 1), was recognized as Mandarin /a/.

The English vowel /ɒ/ was deemed moderately similar to Mandarin /a/, while English /ʌ/ was deemed slightly less similar to the same English vowel. In the L2 production data, both vowels were overwhelmingly recognized as the Mandarin /a/ category. This was predicted for English /ɒ/, but to a lesser extent for English /ʌ/.

The English vowels /ɪ/ and /ɛ/ were deemed to be 'new' and predicted to be more often confused with other English categories rather than any Mandarin category. This prediction was born out in the production data, but in slightly different patterns for responses to each voice used in the elicitation stimuli. In response to Voice 1, /ɪ/ productions were more often recognized as /ɛ/ than the reverse. In response to Voice 2, the opposite was true; /ɛ/ productions were more often recognized as /ɪ/ than the reverse.

Finally, English /u/ was deemed to be 'new' as well as distinct from other English categories. As such, it was predicted to rarely be confused with any Mandarin category. This prediction was not born out. English /u/ was frequently recognized as Mandarin /ɤ/ and to an even greater extent, as Mandarin /u/. Confusions between these categories were most common for productions in response to Voice 1 stimuli. In response to Voice 2 stimuli, more correct productions of English /u/ are evident, although productions were still overwhelmingly in favour of Mandarin /ɤ/ and Mandarin /u/.

Comparing absolute differences across voices and time, Table 5.9 summarizes percentages of L2 English productions that were identified by the Metamodel as the intended English vowel. In addition, percentages of L1 English productions that were identified by the Metamodel as the intended vowel (see Table 3.5 in Chapter 3) are provided as a baseline. It is important to consider L2 production rates in relation to this English L1 baseline data from the Metamodel, rather than to 100% correct targets. Comparing L2 productions of similar vowels to perfect identification as the intended English category would be misleading, since even perfectly acceptable productions of L1 English vowels were often classified by the Metamodel as members of the similar L2 category. For example, although L1 English productions of /i/ were often classified as Mandarin /i/ by the Metamodel, such productions are not necessarily poor members of the English category in absolute terms.

Table 5.9. Percentage of L1 English productions recognized as the intended English category by the Metamodel compared with L2 productions classified by the Metamodel as the intended English vowel by voice and time.

	L1 English	*L2 responses to Voice 1*			*L2 responses to Voice 2*		
Vowel	Baseline	Time 1	Time 2	Difference between times	Time 1	Time 2	Difference between times
/i/	65	33.0	40.9	7.9	39.8	46.6	6.8
/ɪ/	85	54.5	69.3	14.8	83.0	84.1	1.1
/e/	85	21.6	21.6	0	33.0	22.7	-10.3
/ɛ/	82.5	77.3	75.0	-2.3	70.5	76.1	5.6
/æ/	77.5	31.8	55.7	23.9	22.7	46.6	23.9
/ɒ/	60	19.3	28.4	9.1	20.5	30.7	10.2
/ʌ/	61.5	12.5	22.7	10.2	15.9	13.6	-2.3
/o/	60	19.3	19.3	0	18.2	15.9	-2.3
/ʊ/	67.5	15.9	19.3	3.4	11.4	18.2	6.8
/u/	95	18.2	14.8	-3.4	25.0	28.4	3.4

Comparing improvements in production recognition scores over time relative to the English baseline, the results in Table 5.7 suggest that productions of three of the English vowels that are most similar to single Mandarin counterparts, /e/ and /o/ and /ʊ/, are not only rarely recognized as the intended English category, there is no evidence of improvement over time toward a more English-like production. Results of improvement for these vowels generally support the SLM claim regarding L2 learners' inability to discriminate small phonetic differences between similar L1 and L2 categories. However, productions of the similar English vowel /i/ may violate this claim to some degree. Although L2 English productions of /i/ are most often identified as Mandarin /i/, as predicted, L2 production scores demonstrate considerable improvement over time toward higher identification rates as English /i/ in response to both voices, suggesting that some difference is discernable between the L1 and L2 /i/ categories.

The other most similar English vowels /ɒ/ and /ʌ/ both demonstrate improvement in production over time, although only /ɒ/ improves for Voice 2. Given both vowels' relatively high degree of similarity with Mandarin /a/, this is also unexpected.

Comparing differences in production scores over time for the less similar English vowel /æ/, improvement is evident in response to both voices. Improvements on /ɪ/ and /ɛ/ are much smaller, and only in response to Voice 1 for /ɪ/ and Voice 2 for /ɛ/. However, both of these vowels were already highly accurate at Time 1, so less improvement might be expected. Finally, English /u/, which was determined to be a 'new' vowel, violates SLM predictions, demonstrating no clear improvement over time, despite being poorly recognized at Time 1.

Absolute distances from target English categories

While relative classifications of learner productions by the Metamodel based on APP scores provides a great deal of insight, examining absolute statistical distances between L2 productions of a particular category and L1 English productions of the same category provide additional information. Mahalobonis Distance (MD) scores for those L2 English productions that were correctly identified by the English Model as the intended English vowel, regardless of whether they were actually more Mandarin-like in terms of their evaluation by the Metamodel are provided in Table 5.8 below.

Table 5.10. Mean Mahalanobis Distance scores for NS English and L2 English production tokens that were correctly identified by the English Model.

	Mean Score *Std. Deviation* (Number of correctly identified tokens)		
Vowel	NS English Productions (max = 40)	L2 Production Time 1 (max = 176)	L2 Production Time 2 (max = 176)
/i/	6.88	7.80	8.32
	4.19	*4.59*	*6.11*
	(38)	(174)	(175)
/ɪ/	6.68	18.88	20.16
	4.98	*24.36*	*20.08*
	(35)	(119)	(136)
/e/	5.38	14.98	16.93
	3.02	*13.84*	*14.53*
	(39)	(169)	(163)
/ɛ/	5.25	7.37	8.48
	3.18	*4.29*	*5.74*
	(32)	(129)	(128)
/æ/	5.85	12.11	9.31
	4.56	*15.99*	*6.13*
	(34)	(63)	(93)
/ɒ/	7.42	13.69	11.39
	3.95	*14.05*	*10.17*
	(29)	(144)	(162)
/ʌ/	3.45	7.92	9.14
	1.98	*5.18*	*6.12*
	(24)	(51)	(62)
/o/	12.91	17.71	14.64
	13.89	*12.51*	*7.24*
	(39)	(153)	(155)
/ʊ/	6.92	24.99	21.04
	6.47	*17.80*	*15.66*
	(36)	(149)	(155)
/u/	10.75	25.18	22.80
	6.70	*11.50*	*10.04*
	(39)	(51)	(65)
Mean	7.149	15.06	14.22

Mean MD scores for correctly identified L1 English productions are provided as a baseline. This information provides an indication of how well the L2 productions fit the L1 category independent of relation to other categories.

In absolute terms, it appears that degree of crosslinguistic similarity does not accurately predict how well an accented production ultimately fits the intended category. The more 'similar' categories (as defined by the number of production tokens that can be statistically recognized as members of a competing language category) are not necessarily produced in a less English-like fashion than are less similar categories. For example, L2 English productions of /i/ and /o/, similar English vowels, are nearly as close to the English category centres as are L1 English productions of those vowels, while L2 productions of English /e/ and /ʊ/ are relatively poor members of the intended category compared to their L1 English counterparts. For the less similar English categories, distance from the English category seems to vary, with some being closer to the English categorical centre (e.g., /ɛ/, /æ/, /ʌ/ and /ɒ/) than others (e.g., /ɪ/ and /u/).

5.4. Discussion

5.4.1 Identification Confusion Patterns

The confusion patterns for the Lengthened Vowel, Natural Vowel and Generalization identification tests largely support PAM's predictions. When L1 Mandarin categories are very similar to L2 English categories, the learners appear to bootstrap on those L1 categories, resulting in near ceiling identification rates. One minor exception was English /ʊ/, which was sometimes confused with English /ʌ/ or /u/.

Although the general confusion patterns for the Lengthened Vowel, Natural Vowel and Generalization test stimuli are similar, regardless of the test stimuli, interesting differences are also evident. First, the overall identification rates were higher for the Generalization test stimuli. This likely reflects the fact that these stimuli were extracted from the production elicitation stimuli and, in that context, were carefully articulated by phonetically sophisticated speakers. As such, these productions were likely clearer than the stimuli used in the Lengthened Vowel and Natural Vowel tests, thereby resulting in higher identification rates overall; identification stimuli used in the other tests were produced by naïve L1 English speakers in response to auditory elicitation stimuli and variation was greater. Indeed, as was indicated in Chapter 4 (also see Appendix 3, Tables A3.1-A3.4), 100% of the /b, pV/ test stimuli for both voices were correctly recognized by the English Model when it included vowel duration as a variable; 97.5% of productions were correctly identified when duration was excluded (Voice 2 productions of /bæ/ and /pɒ/ were not recognized as intended). In contrast, only 91% of the stimuli used for the Natural Vowel test were recognized as the intended vowel by the English Model when it included vowel duration as a variable; only 86% were recognized as the intended vowel when duration was excluded (refer to Table 3.5 in Chapter 3 for full details).

A second interesting difference in relation to test type is that confusion patterns for /i/-/ɪ/, /æ/-/ɛ/, /ɒ/-/ʌ/ demonstrated higher identification rates on the first vowels in these pairs on the Lengthened Vowel test than on the Natural Vowel test. On the Natural Vowel test, identification rates for the second vowels in these pairs were higher than they identification rates on the Lengthened Vowel test. This finding suggests some learners may be making non-nativelike use of vowel duration for identifying vowels in these contrasts. On the Lengthened Vowel test, they were more likely to perceive some vowel tokens in those pairs as the longer member of the pair than on the Natural Vowel test, where they were more likely to perceive some vowel tokens to be the shorter member of the contrast. It should be noted that, although relative vowel duration may serve as an additional cue to identification and may be useful in the context of otherwise ambiguous stimuli, a number of learners in this study had clearly been informed that this was a primary cue to vowel identification in English. This was reflected by the fact that many of the participants from across training groups complained that for many tokens, they could not accurately judge relative vowel duration, and apologetically emphasized that this was the main source of their errors.

Interesting differences in confusion patterns also emerge when looking at differences between responses to Voice 1 versus Voice 2 on the Generalization test. For a number of vowels, identification rates are clearly different, depending on which voice was heard as the stimulus. This suggests that although each voice produced tokens that were carefully articulated and readily identifiable to the researcher as members of the target category, L2 learners' identification rates for each token reflect specific interactions with the learners' pre-existing categories. L2 learners appear to be very sensitive to how specific tokens of some vowels interact with L1 (and perhaps L2) category boundaries.

5.4.2. Production Confusion Patterns

The general confusion patterns for L2 production data largely support the predictions of the SLM. When L1 Mandarin categories are very similar to L2 English categories, in production, learners tend to substitute the L1 category for the L2 category. One exception to this is English /ʊ/, which although determined to be very similar to Mandarin /ɤ/, was frequently substituted with a vowel that was recognized as more like Mandarin /uə/ in L2 production.

Productions of those L2 categories that are less similar to Mandarin categories were nearly always recognized as either the intended L2 category, or as a member of another L2 category; rarely were they recognized as a member of an L1 category. In other words, learners appear to discern that these are 'new' categories, although incomplete development of contrasts with other L2 categories may limit learners' initial accuracy in production – the learners representation of the 'new' vowels may be more variable at early stages of development. If learners did not discern these to be 'new,' or at least a bad fit to any existing Mandarin category, we should expect more production tokens of these categories to be recognized as a Mandarin category by the Metamodel. The one exception is English /u/. Although this vowel was determined to be unlike any Mandarin category, it was more often recognized as Mandarin /u/ or Mandarin /ɤ/, than as the intended or any other English category.

The SLM offers a potential explanation for the finding that the 'similar' English vowel /ʊ/ and 'new' vowel /u/ were not recognized as predicted. Flege et al.(2003) argue that when adults acquire some L2 categories that are close to existing L1 categories, dissimilation occurs, whereby both the L1 and L2 category boundaries move away from each other so as to prevent potential confusion between the two. This has the ultimate effect of rendering both categories slightly dissimilar from the intended L1 or L2 categories. It is possible that in the process of attempting to dissimilate English /ʊ/ and English /u/ from nearby categories, general confusion ensues. That area of the Mandarin vowel space appears to contain a large number of L1 and L2 categories. The results of testing the English L1 vowel against the Mandarin Model pattern recognition model in Chapter 3 (Table 3.8) illustrate this. In that approach, productions of English /u/ were tested against a pattern recognition model trained solely on Mandarin L1 vowel data. The results indicate that the Mandarin Model recognized 42.5% of English L1 /u/ productions as Mandarin /ɤ/, 40% as Mandarin /u/ and 17.5% as Mandarin /o/. This indicates that three Mandarin vowels are in the vicinity of English /u/, suggesting that this portion of the Mandarin vowel space is particularly dense. Confusions among these vowels may be the result of attempting to integrate the L2 English /u/ category into an already crowded perceptual space.

Another possible explanation for the failure to predict this observed behaviour is that the productions were elicited by only two voices, neither one of which may have been ideal stimuli for this group of learners.

In addition to successfully predicting most Metamodel confusion patterns, partial support for the SLM is also provided in the results evaluating improvement on specific vowels from Time 1 to Time 2. It appears that for most English vowels, the degree of similarity between L1 and L2 categories predicts improvement. Three of the four most similar vowels did not improve over time, while those that were least similar, with the exception of /u/, either improved, or if they did not, were already relatively accurate at Time 1 and therefore had less room to improve.

Finally, Mahalanobis Distance (MD) scores appear to contradict SLM claims that L2 vowels that are least similar to L1 categories will be produced by L2 learners as better members of the L2 categories than those L2 vowels that are more similar. The nature of the production task, through elicited imitation using sentence frames, was intended to ensure that the speech stimuli were actually categorized and reproduced in terms of the learners' phonological systems, rather than through oral mimicry. If categorization in terms of the learner's phonological system occurs, then production values for the L2 category should reflect that system. The success of this approach is demonstrated by the many instances where a similar L1 vowel was seemingly substituted for the L2 category. The MD scores for some L2 productions of less similar English vowels suggest that learners' representations of those vowels can be far from accurate, indicating a possible inability to discern small phonetic differences within the new L2 category. An alternative explanation for this phenomenon is that although differences are discernable, L2 category formation reflects the process of dissimilation proposed by the SLM. It may be that because the learner is trying to fit more categories into a finite perceptual space, the average location of even 'new' categories may be shifted to better accommodate other L1 and L2 categories. This possibility needs to be explored in a further study. If the process of dissimilation is common, it predicts that any improvements in MD scores for one category might come at the expense of MD scores for another category.

5.4.3. Relationship between perception and production

There are several striking similarities between the L2 learner identification data and the L2 production data indicating a strong connection between L2 perception and production. In general terms, similar L2 vowels were identified with a high degree of accuracy by apparently bootstrapping on L1 categories. This was reflected in L2 production, where the same vowels were overwhelmingly recognized by the Metamodel as members of the L1 category.

Comparing L2 learner identification response patterns for the less similar vowels, by voice, provides the strongest indication of a direct connection between perception and production. In the identification task, English /æ/ was frequently misidentified by L2 listeners as English /ɒ/, for Voice 1 only. This would most likely be the case if productions of English /æ/ by Voice 1 were more similar to the Mandarin /a/ category than were productions of English /æ/ produced by Voice 2. Since English /ɒ/ largely bootstraps on Mandarin /a/, if a Mandarin learner of English perceives Voice 1 /æ/ productions to be similar to Mandarin /a/, this might result in confusion with English /ɒ/ on the identification test. In fact, this is precisely the case. When the production elicitation CV stimuli were tested against the Metamodel in the previous chapter (also refer to Appendix 3, Tables A3.9 and A3.10), the APP scores on Mandarin /a/ for the

English vowels in the /bæ/ and /pæ/ stimuli produced by Voice 1 were found to be 0.09 and 0.37 respectively. In contrast, the APPs that the two productions of English /æ/ by Voice 2 were more like Mandarin /a/ were found to be 0.00 and 0.01 respectively. This suggest that on the identification task, some learners perceived Voice 1 /æ/ to be like Mandarin /a/ and therefore identified it as English /ɒ/, which is also similar to Mandarin /a/. If this analysis is correct in describing the assimilation process that occurred during the identification test, we should expect that L2 English productions in response to Voice 1 /æ/ should more often be identified by the Metamodel as a member of Mandarin /a/ than L2 productions in response to Voice 2 /æ/. This is exactly what occurred. A large number of L2 productions of /æ/ in response to the Voice 1 stimuli were recognized as Mandarin /a/, while virtually no productions in response to Voice 2 were recognized as that Mandarin category. Instead, confusions in productions of /æ/ in response to Voice 2 stimuli were primarily found to be with English /ɛ/.

Another clear difference in the identification rates for vowels varying by stimulus voice was for the English vowel /ɛ/. It was found that on the identification test, the Voice 2 /ɛ/ stimuli were misperceived as /ɪ/ far more often than was the case for the Voice 1 stimuli. Again, this pattern was also demonstrated in production. L2 productions of /ɛ/ in response to Voice 2 were far more frequently recognized as /ɪ/ than in response to Voice 1. The English Model's recognition of the /bɛ/ and /pɛ/ stimuli varied across stimulus voice. While both Voice 1 and Voice 2 productions of this vowel were accurately recognized by the English Model in absolute terms, productions by Voice 2 had a slightly higher a posteriori probability of being English /ɪ/ than did productions by Voice 1 (refer to Table A.3.1 and A3.2 in Appendix 3).

These similarities in L2 learner identification and production patterns demonstrate that while L2 learners are heavily influenced by L1 categories when the L2 input is similar to those categories, they are still able to discern small phonetic differences between two native speakers' productions of the same 'new' L2 category, as was the case in learner responses to English /ɛ/, varying by stimulus voice. Put differently, native speaker variation in the production of similar L2 vowels (e.g., /i/) does not seem to affect ultimate identification and production by L2 learners. Perhaps the strength of assimilation by the L1 category may wash out any differences between voices in much the same way differences between speakers are normalized in L1 perception. In contrast, for less similar L2 vowels (e.g., /ɛ/), small differences in stimulus voice matter and are used to form new categorical boundaries.

Chapter 6. A brief comparison of training with naturalistic L2 vowel development

The purpose of this chapter is to briefly compare the results of the Mandarin vowel training study described in Chapters 4 and 5 with data from an earlier longitudinal study of naturalistic L2 English vowel learning. In this earlier study, Mandarin L1 speakers were tested six times over a ten-month period to determine to what extent their ability to produce English vowel contrasts developed in the absence of explicit instruction. The L2 production data from this naturalistic vowel learning study have been previously evaluated by human listeners in Munro et al. (2003) and Munro and Derwing (2007). In addition, Thomson (2005) reports a preliminary attempt at statistical evaluation of these data. The production elicitation methodology used for the data collection in the naturalistic vowel learning study was similar, although not identical, to that used for the training study reported in Chapters 4 and 5 of this dissertation. The ten English vowel categories being evaluated were identical. While the earlier data have the potential to provide insight into English vowel learning in the absence of instruction, differences between the samples in each study and in the methods of data collection impose limitations on the conclusions that can be drawn.

6.1. Research Questions

The comparison of naturalistic L2 English vowel learning data with the L2 English vowel learning data from the training study that is the focus of this dissertation is intended to address the following three research questions:

1. To what extent do English L2 vowel productions by Mandarin L1 speakers improve over the course of ten months in the absence of focused training in L2 speech perception?
2. Is improvement in English vowel production for the trained group more rapid than it is for the naturalistic learning group?
3. Are relative difficulties associated with acquisition of each L2 English vowel category similar in naturalistic versus trained English vowel learning contexts?

6.2. Method

6.2.1. Speakers

Native English speakers

L1 English vowel production data were obtained from 33 undergraduate students at the University of Alberta (7 males, 26 females; *M* = 25.85, range = 19 – 44). All were either born in Alberta or had arrived at a very young age and had spent most of their entire lives there. In addition, while several reported advanced knowledge of a second language, all used English as their primary language. None reported advanced knowledge of Mandarin. All reported normal hearing.

L2 English speakers

The L2 English vowel production data for the naturalistic vowel learning study were obtained from 20 native Mandarin speakers (6 male, 14 female; ages 26-42, *M* = 33.2). The participants were well-educated immigrants to Canada who had arrived in the country less than four months (*M* = 2.8 months, range = 1- 4 months) before participating in this research. All had started a fulltime ESL program approximately one month prior to the first data collection and reported having minimal exposure to English prior to beginning their formal ESL studies. They were assessed as being between Benchmarks 1 & 3 on listening and speaking skills (Stage 1), according to the Canadian Language Benchmarks Assessment tool (Smith, 2000). This means they were at a beginner proficiency level, although some had studied English in China and were able to read and write at higher Benchmark levels. The participants' English vowel production ability was tested and recorded on six occasions at two-month intervals. Most were still taking ESL at the time of the sixth recording, ten months after the first recording. The ESL program included no focused instruction in English pronunciation. Finally, each participant passed a pure-tone hearing test at the onset of the study.

6.2.2. Procedure

Participants from both the L1 English and L2 English groups were tested individually in a quiet room and recorded using a high quality digital recorder with a sampling rate of 44,110 Hz. For the L2 English learners, this data collection was only one part of a battery of tests conducted at the same time as part of a larger study of L2 speech development. As previously stated, recordings of the L2 participants' English vowel productions were made on six occasions, at two-month intervals.

The elicited production stimuli used for the L1 and L2 English speakers in the naturalistic vowel learning study were the same. Participants were asked to listen to and repeat a series of /bVt/ and /pVt/ stimuli containing the target vowels. One exception was made, whereby /bʊk/ was used instead of /bʊt/ because it was a real word that was likely to be readily identifiable by learners. The entire set of target words is provided in Table 6.1. With the exception of /put/, all are real words, although their frequency and recognizability as real words varies. The auditory stimuli were spoken by a female native speaker from Edmonton, Alberta, the same speaker who provided the Voice 1 stimuli for the training study reported in Chapters 4 and 5.

Table 6.1. Training and testing stimuli for ten English vowels.

	English /bV/ targets									
IPA	bit	bɪt	bet	bɛt	bæt	bɒt	bʌt	bot	bʊk	but
	English /pV/ targets									
IPA	pit	pɪt	pet	pɛt	pæt	pɒt	pʌt	pot	pʊt	put

All CVC stimuli were presented in the carrier phrase, "*The next word is* _____" and participants were asked to respond by repeating the word in the carrier, "*Now I say* _____". The entire procedure was conducted once for each participant. After recording each of the speakers' productions, the target syllables were extracted from the sentence frame. Next, they were down sampled to 22.055 kHz, normalized across tokens to peak amplitude, and saved in separate sound files for each word.

6.2.3. Data analysis

The same approach to vowel measurement used for the L1 production data described in Chapter 3 and the L2 production data described in Chapter 4 was used to evaluate the English L1 speakers' CVC productions and the English L2 speakers' CVC productions in this study. Using the same acoustic analysis software, formant frequency and pitch measurements as well as vowel duration were extracted for evaluation. For the L1 English CVC data, due to an inability to satisfactorily extract acceptable formant and/or pitch measures, 12 of 660 production tokens were not analyzed. For the L2 English CVC data, due to either missing recordings or an inability to satisfactorily extract acceptable formant and pitch measures, 24 of the total 2400 production tokens were not analyzed (four tokens from Time 1; six at Time 2; one at Time 3; four at Time 4; four at Time 5 and six at Time 6). For missing cases, the same participants' values from their production of the same vowel in the alternate CVC context were substituted.

Values for F1, F2 and F3 taken from the 20% and 70% points of each L1 English vowel token's length, as well as F0 and duration were used to train and test an English pattern recognition model for these vowels. Since the model was being trained and tested on the same production data, I used a round-robin cross-validation approach whereby each speaker to be tested was excluded from the training set on which his or her productions would then be tested. After establishing that the model was reasonably accurate in its ability to classify the L1 English data, values for F1, F2 and F3 taken from the 20% and 70% points of each L2 English vowel token's length as well as F0 and duration were then tested against the English CVC model. Using further statistical tests, improvement in the learners' performance over time was measured.[20]

6.3. Results

6.3.1. Naturalistic L2 English learners' vowel production data

The results illustrated in Table 6.2 below demonstrate that the English CVC Model, including duration as a variable, was very accurate in recognizing L1 English CVC productions; 93.9% of cases were recognized as the intended vowel. As with the English CV Model in Chapter 3, some variation in recognition rates across vowel categories is evident.

---

[20] The results reported in this chapter reflect tests of L2 data against the CVC English Model and CV English Model which include duration as a variable. Alternate results which exclude vowel duration as a variable in the pattern recognition models are provided in Appendix 9. Despite the fact that excluding vowel duration results in slightly lower mean correct recognition scores, the overall results concerning which factors are significant are the same.

Table 6.2. Recognition of English CVC production tokens by vowel tested against the CVC English Model trained and tested on native speaker English productions with vowel duration included as a variable. Values represent percentages of intended vowels recognized as belonging to each English vowel category.

		Vowel recognized by CVC English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended English vowels repeated in response to auditory stimuli	/i/	**95.5**	--	4.5	--	--	--	--	--	--	--
	/ɪ/	--	**95.3**	--	4.7	--	--	--	--	--	--
	/e/	4.5	--	**95.5**	--	--	--	--	--	--	--
	/ɛ/	--	--	--	**98.4**	1.6	--	--	--	--	--
	/æ/	--	--	--	3.1	**95.3**	--	1.6	--	--	--
	/ɒ/	--	--	--	--	--	**93.8**	6.2	--	--	--
	/ʌ/	--	--	--	--	--	3.2	**91.9**	--	4.8	--
	/o/	--	--	--	--	--	--	--	**98.5**	--	1.5
	/ʊ/	--	--	--	--	--	3.1	4.7	1.6	**89.1**	1.6
	/u/	--	--		--	--	--	--	--	3.0	**95.5**
Total correct		94.9% (93.8% without vowel duration cue)									

Having established that the English CVC pattern recognition model adequately classifies L1 English productions, mean classification rates of the L2 production data by the English CVC model are analyzed below. Figure 6.1 provides a summary of the learners' mean correct L2 English vowel production scores over time.

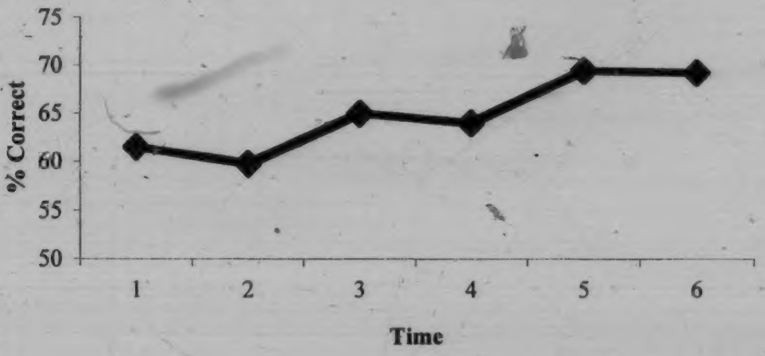


Figure 6.1. Mean correct production scores of untrained speakers' L2 English vowel productions as identified by the English CVC Model (without vowel duration).

A two-way repeated measures ANOVA with Time (6 levels) and Vowel (10 levels) as within-group factors demonstrated a significant difference for Time [$F(5,95) = 3.976, p = .003$] and Vowel [$F(9,171) = 15.603, p < .000$]. No significant Time x Vowel interaction was found[21]. Follow-up Bonferroni-adjusted paired samples *t*-tests and Tukey HSD tests were conducted to determine if improvement could be detected between consecutive data collection points (i.e., between each two month interval). No significant differences between any two consecutive data collection points were found. Further Bonferroni-adjusted paired samples *t*-tests indicated that the only significant differences between data collection points were between Times 1 and 5, Times 1 and 6 and Times 2 and 6.

The absolute mean correct recognition scores for L2 productions at each time indicate that the mean difference between Times 2 and 3 (5.25%) and between Times 4 and 5 (5.5%) accounted for all of the improvement detected over time. Between Time 5 and 6, no further improvement is indicated by the mean correct recognition scores.

6.3.2. Naturalistic vowel learning versus instructed vowel learning

The following comparison of the naturalistic English vowel learning group with the training study group is tentative. Differences between the participants and methods used in the naturalistic vowel learning study and the vowel training study introduce potentially confounding variables. For example, by Time 6, participants in the naturalistic vowel learning study had been in Canada for between 11 and 14 months. Most of that period had been spent in an ESL classroom. In contrast, the participants in my training study had been in Canada for between 4 and 48 months, although only four of the 22 had lived in Canada for more than a year. In terms of LOR in Canada, then, the naturalistic group at Time 6 and the Training group at Time 1 and 2 are relatively comparable. However, a larger difference is found when comparing each group's prior English training and ability. Nearly all the participants in my training study had been attending ESL classes for less than six months at the end of the study. Additionally, although the participants in both studies entered ESL training at beginner proficiency levels, in general, the naturalistic study participants had slightly lower proficiency levels overall. Many were at CLB Benchmarks 1 and 2 when they began to participate in the study. In contrast, most of the learners in my training study began ESL training at CLB Benchmark 3. Finally, the production stimuli in the two studies were different; CVC targets were used for the naturalistic vowel learning study, while CV targets were used for the training study.

---

[21] Mauchly's test of Sphericity was not significant, however, results for corrected and multivariate tests are provided in Appendix 8, Table A8.1.

Table 6.3. Mean percent correctly recognized vowel productions over time for the naturalistic group's L2 productions (top panel), in contrast to the trained group's L2 productions described in Chapters 4 and 5 (bottom panel). Vowel duration was included as a variable in both the CVC English and CV English Models.

	Naturalistic learners' /b,pVt/ production data (n= 20)	
Time	% Correct *(SD)*	% Improvement from previous time *(SD)*
1	61.50 *(11.48)*	
2	59.75 *(13.62)*	-1.75 *(12.28)*
3	65.00 *(12.89)*	5.25 *(9.52)*
4	64.00 *(11.19)*	-1.00 *(11.42)*
5	69.50 *(13.95)*	5.50 *(12.13)*
6	69.25 *(10.79)*	-0.25 *(14.55)*

	Trained learners' /b, pV/ production data in response to Voice 1 (n=22)	
Time	% Correct *(SD)*	% Improvement from previous time *(SD)*
1	73.18 *(10.94)*	
2	77.84 *(9.11)*	4.66 *(9.01)*

A comparison of mean percent correct recognition of L2 productions across studies[22] (refer to Table 6.3 above) suggests that the learners in the training study may have already been better able to produce English vowels at Time 1 than the learners in the naturalistic group Time 6. However, the magnitude of improvement for the training group after just three weeks (4.66 percentage points) was nearly as large as the largest improvement over any two-month period in the naturalistic vowel learning study (5.50 percentage points). Dividing this largest improvement by learners in the naturalistic vowel learning study into weeks, an average improvement of 0.69 percentage points per week is evident. This means that over three weeks, the same period during which learners in the training study were tested, learners in the naturalistic study improved an average of 2.06 percentage points. Finally, the naturalistic vowel learning group's rate of improvement in English vowel production appeared to slow (for a second time) after reaching 69.5% accuracy at Time 5. In contrast, the training group's initial performance was already 67% at the beginning of that study, and the rate of improvement during three weeks of training was relatively rapid.

Recognition scores for L2 English vowel productions, by vowel category, for the naturalistic vowel learners and the training study learners are provided in Table 6.4.[23] The mean percent correct recognition scores for each intended vowel production in each study show similar patterns. For example, in both studies, productions of the English vowels /i/, /e/, /ɒ/ and /o/ were more often recognized as the intended vowels than were productions of the remaining English vowels. However, in the naturalistic study, at Time 1, mean recognition scores for some of these vowels were relatively low, and mean recognition scores for /æ/ and /ɛ/ were initially just as high. Furthermore, among the four most accurate vowels, recognition of /i/, /e/ and /ɒ/ productions accounted for 51% of the naturalistic learners' total improvement. In contrast, for the same vowels, only recognition of /ɒ/ productions improved for learners in the vowel training study, and accounted for only 16% of that group's total improvement. The fact that recognition

---

[22] Only production results in response to Voice 1 stimuli from the training study are compared with the naturalistic vowel learning results since the same stimulus voice was used in both studies.

[23] Full confusion matrixes for L2 productions from the naturalistic study and for L2 productions in response to Voice 1 /b, pV/ stimuli in the training study are provided in Appendix 8, Tables A8.2 – A8.9.

Table 6.4. Summary of mean percent correctly recognized L2 English productions for each English vowel, contrasting naturalistic vowel learning results (from Times 1-6) with trained vowel learning results (from Times 1-2). Vowel duration included as a variable.

	Naturalistic Vowel Learning Study (n=20)							Vowel Training Study (n=22)		
Vowel	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Difference from Time 1 to 6 (approx. 10 months)	Time 1	Time 2	Difference from Time 1 to 2 (approx. 3 weeks)
/i/	75	72.5	92.5	77.5	87.5	90	15	97.7	96.6	-1.1
/ɪ/	25	25	27.5	35	37.5	42.5	17.5	51.1	71.6	20.5
/e/	67.5	72.5	82.5	90	92.5	92.5	25	96.6	95.5	-1.1
/ɛ/	65	65	67.5	60	60	55	-10	73.9	75.0	1.1
/æ/	75	72.5	75	87.5	85	87.5	12.5	68.2	78.4	10.2
/ɒ/	70	75	77.5	87.5	85	90	20	88.6	96.6	8.0
/ʌ/	67.5	37.5	52.5	35	52.5	50	-17.5	58.0	63.6	5.7
/o/	92.5	87.5	95	92.5	95	92.5	0	90.9	90.9	0.0
/ʊ/	47.5	40	42.5	37.5	57.5	35	-12.5	76.1	80.7	4.5
/u/	30	50	37.5	37.5	42.5	57.5	27.5	30.7	29.5	-1.1
Mean	61.5	59.75	65	64	69.5	69.25	7.75	73.2	77.8	4.7

of /i/ and /e/ productions in the vowel training study did not improve over time can be explained by the fact that productions of these vowels already had near ceiling recognition scores at Time 1, while in the naturalistic learning study, they did not.

In the naturalistic vowel learning study, productions of /ɪ/, /ʊ/ and /u/ all had the lowest recognition scores at the beginning of the study and the recognition scores for these vowels remained relatively low at Time 6. Among these weakest vowels, productions of /ɪ/ and /u/ in the naturalistic vowel learning study improved substantially, while the mean recognition score for /ʊ/ productions actually worsened. In contrast, in the training study, recognition scores for /ʊ/ productions were initially quite high, and also demonstrated improvement after training. Recognition scores for /ɪ/ and /u/ productions in the training study were relatively weak initially, but only recognition of /ɪ/ productions demonstrated improvement after training.

Productions of the remaining vowels, /ɛ/, /æ/ and /ʌ/, had moderate recognition scores at Time 1 in both studies. However, for the naturalistic learners, only productions of /æ/ were more accurately recognized at Time 6, while productions of /ɛ/ and /ʌ/ were far more poorly recognized. In contrast, in the training study, recognition scores for productions of all three vowels demonstrated improvement.

The most obvious difference between learners in the naturalistic learning study versus the training study, in terms of the relative difficulty each group experienced with particular vowels, is evident in their productions of /ʊ/ and /u/. The learners in the training study performed much better on /ʊ/ than the learners in the naturalistic study. Conversely, learners in the naturalistic study performed much better on /u/ than did the learners from the training study. However, the nature of the statistical production recognition confusions for each of these vowels by each group differs[24]. For the naturalistic learners, /ʊ/ productions were often recognized by the English Model as /o/, while for the training group, /ʊ/ productions sometimes, but infrequently, were recognized as /ʌ/ or /ɒ/. For learners in the training group, /u/ productions were frequently recognized as both /o/ and /ʊ/, while for the naturalistic learners, /u/ productions were primarily confused with /o/. These patterns suggest that differences in performance on these vowels may be related to the perceptual properties associated with each vowel stimulus, rather than necessarily indicating differences in the learners' ability to discern these vowels.

---

[24] Tables A8.2 – A8.9 in Appendix 8 provide full confusion matrixes.

In summary, in the naturalistic vowel learning study, the CVC English Model indicated improvement in the learners' production of six vowels, /i/, /ɪ/, /e/, /æ/, /ɒ/ and /u/. However, 51% of the total improvement was accounted for by productions of three vowels, /i/, /e/, and /ɒ/, productions of which were relatively well produced even at the beginning of the study. In contrast, for the training study, the CV English Model indicated that although learners also demonstrated mean improvement in six vowels, /ɪ/, /ɛ/, /æ/, /ɒ/, /ʌ/ and /ʊ/, all but /ɒ/ were relatively weak to begin with. Productions of the more accurately produced vowels could not improve very much, because their recognition scores were already near ceiling before training began. In contrast, improvements in the five more difficult vowels accounted for 86% of the training group's total improvement.

6.4. Discussion

At the beginning of this chapter I posed three research questions. First, I asked to what extent English L2 vowel productions by Mandarin L1 speakers improve over the course of ten months in the absence of focused training in L2 speech perception. The results of this study indicate that while significant improvement occurred, the rate of improvement was very slow. The general results reported here are very similar to those found by Munro et al. (2003) and Munro and Derwing (2007) where human listeners were asked to categorize the same L2 productions used in this study – the learners were found to have improved over time to the same extent. The pattern recognition model's recognition of the learners' productions of specific vowels also appears to reflect the human listener responses reported in the earlier studies (Munro et al., 2003; Munro & Derwing, 2007). The relative difficulty the learners experienced with particular vowels was very similar. Furthermore, the relative improvement found within specific categories over time was very similar. This provides preliminary evidence that the statistical model employed here roughly approximates human listener judgments. The fact that the results in this study are not identical to those reported in Munro and Derwing (2007) is not surprising given the fact that even the four listeners used in their study demonstrated complete agreement on only 67% of the items they heard.

My second research question asked whether improvement in English vowel production for the participants in the training study was more rapid than for the naturalistic learning group. The answer is yes. While the largest absolute improvement in the naturalistic vowel learning group over a two-month period was an increase of 5.5%, in the training group, improvement after just three weeks was 4.66% for responses to the same voice used to elicit productions in the naturalistic study. Furthermore, nearly half of the improvement in the naturalistic group's production recognition scores occurred within three categories that are considered the most 'similar' to Mandarin categories (i.e., /i/, /e/, /ɒ/). In the training study, productions of these vowels were already quite accurately recognized in L2 production even before training began. Only production recognition scores for /ɒ/, the weakest among them, demonstrated improvement.

One possible explanation for the differences in initial performance on these vowels may be related to differences in the production elicitation stimuli that were used in each study. In the naturalistic study, CVt syllables were presented and elicited. This syllable structure violates Mandarin phonotactic constraints, which disallow coda stop consonants. As such, errors in production in this context may stem from the learners inability to perceive and/or produce the target vowel in this new context. In the production task, they may have been focusing as much on producing the final consonant as they were on producing the vowel. Given that productions of most 'similar' vowels were not accurately recognized at Time 1 in the naturalistic training study, it may be that improvements in the production of these vowels were not due to learning the vowels themselves, but rather, the result of learning to produce familiar vowels in a new context. Conversely, in the training study, productions of these 'similar' vowels (i.e., /i/, /e/, /ɒ/) were initially well recognized as the intended vowel. This may be due to the fact that they were presented in CV syllables that not only do not violate Mandarin phonotactic constraints, but also are real words in Mandarin. As a result, transfer to an English context is easier.

My third research question asked whether relative difficulties associated with acquisition of each L2 English vowel category are similar in naturalistic versus trained English vowel learning contexts. As the response to the previous research question also partially indicated, the answer to this question is yes. The patterns in terms of which vowels were relatively easy to acquire and which were relatively difficult were virtually the same. For the two vowels for which performance was clearly different across the two groups of learners (i.e., /ʊ/ and /u/), examining interactions with other vowel categories suggests that these differences may be the result of differences in the spectral properties of the elicitation stimuli used in each study.

Given the differences in the populations studied and the stimuli used, this comparison of naturalistic L2 English vowel learning with results from the training study reported in Chapters 4 and 5 offers only preliminary insight. It appears to raise important questions about the ability of learners to easily acquire some L2 English vowel contrasts without explicit instruction. These preliminary findings demand further investigation. Research needs to be conducted which compares naturalistic learning with the effect of training using deliberately matched samples and identical stimuli.

Chapter 7. General summary and discussion

The goal of my dissertation research was to expand our current understanding of L2 speech perception and production. First, I proposed a new statistical model for measuring crosslinguistic similarity and used it to make predictions concerning English vowel learning by Mandarin L1 speakers. I then applied this statistical approach to measuring the effect of training on Mandarin L1 speakers' ability to identify and produce ten English vowels. Finally, I briefly compared the effect of training with naturalistic English vowel learning data from an earlier study to determine if any preliminary insights might be gleaned.

This final chapter provides a general summary of my research. I begin by assessing the general efficacy of applying a statistical pattern recognition Metamodel to the problem of L2 speech perception and production. I then review my findings from the training study in terms of general L2 speech learning issues reviewed in Chapter 1 and discuss implications in terms of current models of L2 speech perception and production. I also briefly describe some potential pedagogical implications of this research. Finally, I conclude by providing suggestions for further research.

7.1. Measuring crosslinguistic vowel similarity using the Metamodel

Flege's (1995) SLM and Best's (1995) PAM make a number of explicit predictions regarding L2 learner behaviour based on interactions between L2 learners' L1 phonological systems and the target L2 phonological system. In applying SLM and PAM to L2 learner data, the validity of any claims that are made hinges on researchers' ability to operationalize the construct of crosslinguistic similarity. Flege (2005) stated that the constructs of 'similar' and 'new' L2 sounds remain inadequately defined, and that a more precise form of measurement could potentially lead to new insights in our understanding of L2 speech learning. In recent years, the most commonly used approach to measuring crosslinguistic similarity has been to plot mean F1 and F2 values, taken from vowel production data of two or more languages under examination, in a two-dimensional space. The resulting distributions of the L1 and L2 language categories are then compared with each other, often informally and subjectively. The degree to which competing language category distributions overlap is used to define the degree of crosslinguistic similarity. Morrison (2006), Strange (2007) and Thomson (2005) have recently applied a statistical pattern recognition approach to measuring crosslinguistic similarity. This provides a much more precise means of measuring crosslinguistic similarity than the currently popular two-dimensional approach just described, although clearly it is an extension of it. A pattern recognition approach is more precise because it can incorporate multidimensional cues used in vowel identification, including measures of F1, F2, F3, pitch and duration. Morrison (2006) and Strange (2007) used this approach to determine which L1 categories were closest to which L2 categories by testing production data from one language against a statistical model trained on production data from another language. However, these researchers assumed that at least initially, L2 productions must all assimilate to the nearest L1 category. This has not been shown to be the case. In fact, both the SLM and PAM explicitly acknowledge that some productions in one language may be perceived as poor examples of any category in an opposing language.

In this dissertation, I have extended the pattern recognition model further (following Thomson, 2005), testing production data from each language being compared against a single statistical pattern recognition model (what I have termed the Metamodel) that has been trained on production data from both of the languages being contrasted. The extent to which production tokens from one language are misclassified as a member of a competing language category appears to provide a more precise means of measuring crosslinguistic similarity.

One advantage of the Metamodel approach tested in Chapter 3 is that production tokens in the competing language's categories can be individually specified as either more or less similar to opposing language categories. For example, in the crosslinguistic similarity study reported in Chapter 3, nearly all productions of English /i/, /e/, /ɒ/ and /o/ were recognized as members of Mandarin /i/, /e/, /a/ and /o/ categories by a Mandarin Model trained only on Mandarin production data. In contrast, the Metamodel, trained on both Mandarin and English production data, recognized that some English productions, by certain speakers, were more Mandarin-like than were others and that some Mandarin productions, by certain speakers, were more English-like than were others. Alternately, for some English categories (e.g., /ɪ/ and /ɛ/), very few if any productions were ever classified as a member of a competing Mandarin category. This indicated that the distributions of these English categories, as a whole, are relatively dissimilar from any competing language category.

When Mandarin accented productions of English vowels that are most 'similar' to Mandarin vowels were tested against the Metamodel, the majority of those productions were recognized as being most like the 'similar' Mandarin vowel category. Furthermore, the resulting proportions of correct classifications of those L2 productions as the intended English vowel vis-à-vis the 'similar' Mandarin vowel resembled the proportion of Mandarin L1 vowel productions of the 'similar' Mandarin vowels that were recognized as being most like the 'similar' English category. This relationship suggests that the Mandarin learners are simply producing Mandarin vowels in place of 'similar' English vowels. The ultimate accuracy of the Metamodel predictions regarding learner behavior suggests that this approach is a very effective means of detecting small differences between L1 and L2 categories in general, and more importantly perhaps, this approach is able to detect small differences between individual production tokens within categories.

The Metamodel approach also yields information concerning potential within-language similarity; some L1 vowel production tokens may not be clearly recognized as members of the intended category, but rather, as members of a nearby category within the L1. This can lead to predictions regarding potential developmental difficulties associated with L2 phonological learning of some L2 categories, where within-language confusion may occur for L2 learners. This was shown to be the case for native speaker productions of English /ɪ/ versus /ɛ/; misclassifications of native speaker English /ɪ/ as /ɛ/ as well as the reverse, accurately predicted that L2 learners would also confuse these categories, despite being relatively free of any L1 influences in the acquisition of these 'new' categories.

One limitation of this study is that the sample sizes I used to compare Mandarin and English vowel inventories were relatively small. For a more accurate measurement of crosslinguistic similarity, a larger sample size is needed. This would better represent confusion patterns in the productions of the larger population from which the sample was obtained. Another limitation is that Metamodel comparisons are constrained by the extent to which vowel production data from each language can be obtained in nearly identical phonetic contexts. For example, comparing vowels found in a CV context with those found in a CVC context might yield inaccurate results. Because of this constraint, it is possible that the Metamodel approach to crosslinguistic similarity might miss the potential influence of allophonic variants found in other contexts that are not being studied. For example, in the case of Mandarin, other allophones, found in contexts other than the /b, pV/ contexts chosen for comparison, may provide Mandarin learners of English with a foundation for learning some English vowel categories. However, a statistical pattern recognition model must allow for contextual effects to ensure reliability in classifying new production data. Consequently, the Metamodel cannot easily reflect a human listener's ability to perceive that different productions of the same vowel produced in different phonetic contexts are indeed the same vowel, despite such productions sometimes having slightly different spectral properties. I was able to overcome this obstacle in the case of Mandarin /o/. Although it does not occur in a post-labial context in Mandarin, it does occur in other contexts. As a result, I was able to elicit production of this Mandarin vowel in the post-labial context by having speakers first refer to how the vowel was produced in other contexts.

Finally, the results of the Metamodel cannot easily be validated against human listener data. Given the nature of potential native language magnet effects (Kuhl & Iverson, 1995), we cannot expect that a Mandarin-English bilingual would be able to distinguish between more Mandarin-like versus more English-like productions of similar categories such as English /i/ and Mandarin /i/, or English /e/ and Mandarin /e/. As I stated in Chapter 2, the Metamodel represents an idealized bilingual speaker's interlanguage – a speaker for whom discerning small phonetic differences between similar L1 and L2 categories is possible. Since such a speaker is unlikely to exist in the real world, the Metamodel should be understood for what it is, a statistically sophisticated approach to measuring crosslinguistic similarity, rather than an ontological statement. While the Metamodel appears to work quite well in the context of this study of Mandarin and English vowels, its validity as a measurement tool needs to be confirmed by assessing its ability to predict learner behaviour in a variety of L1/L2 contexts.

7.2. The effect of training on L2 vowel perception and production

7.2.1. Comparing differences in vowel identification training conditions

In Chapter 1, I summarized a variety of perspectives concerning the nature of L2 phonological learning. I began with the assumption that naturalistic phonological learning was fundamentally different for L1 learners and L2 learners. The most obvious difference is that L1 learners begin with a blank slate, while L2 learners already have a phonological system in place that interacts with the new L2 system being developed. For older L2 learners especially, the strength of previously established phonetic categories appears to play a major role in limiting the learners' ultimate ability to speak the L2 without a detectable accent. I stated that while the effect of L1 cannot be removed from the L2 learning process, instructional intervention provides an opportunity to recreate some beneficial features of L1 acquisition. In particular, controlled L2 learning environments afford the opportunity to provide learners with input that is more likely to be noticed and subsequently incorporated into the learners L2 system. The effect of classroom instruction on SLA has largely been limited to grammatical and lexical domains (cf. Doughty & Williams 1998, Lyster and Ranta, 1997; Sharwood Smith, 1993; Van Patten 196, 2002; White, 1998). Less research is available pertaining to the effect of L2 pronunciation instruction. Seminal research by Derwing et al. (1998) has demonstrated that while pronunciation instruction appears to have a positive impact on the acquisition of L2 prosodic features, the impact of instruction on segmental features is less obvious; in Derwing et al.'s study, segmental learning, although evident in a reading context, did not appear to transfer to extemporaneous L2 speech production. Other research by Munro and Derwing (2007) indicates that while naturalistic learning of English vowels does occur, the rate of acquisition is very slow for both Mandarin and Russian English learners. There is also some indication that most improvement in L2 phonetic learning occurs immediately after arrival in the L2 context, and then begins to asymptote far short of nativelike ability (Jia et al., 2006; Munro and Derwing, 2007).

In the training study reported in Chapter 4, I wanted to further explore the impact of instruction on L2 phonetic learning. Recent research has suggested that the ability to discern new L2 phonetic contrasts can be facilitated by orienting learners attention to the phonetic information associated with specific contrasts. Guion and Pederson (2007a) demonstrated that off-line orienting of L2 learners' attention to phonetic versus semantic information had a measurable effect on learning. In review, those learners who were instructed to attend to the phonetic structure of the L2 words they were being trained to identify demonstrated greater improvement in phonetic learning than learners who were instructed to attend to the meaning of the words. In another study, (Guion & Pederson, 2007b) found that orienting learners' attention to the consonants vis-à-vis the vowels in the training stimuli facilitated the learning of novel L2 consonant contrasts; learners who were instructed to attend to the vowels in the same stimuli failed to improve in their ability to discriminate those consonant contrasts.

Although I had hoped that all learners in my study would benefit from training, I also attempted to increase the likelihood that learners would detect important phonetic information by manipulating the training stimuli in two ways. For one group, I modified training stimuli by lengthening the vowel portion of the training syllables, hypothesizing that this might give the learners a longer period during which they would be able to detect important phonetic information; in another condition, training stimuli were selected that were statistically less similar to Mandarin vowel categories than naturally varying English vowel stimuli. I hypothesized that this might allow the learners to detect differences between the target English categories and similar Mandarin categories. Flege's (1995) SLM and Best's (1995) PAM both argue that difficulty in acquiring L2 categories is related to the degree of similarity between L1 and L2 phones. If more of the L2 input comprises tokens that are less Mandarin-like, it seems reasonable to conclude that learners will have a better chance of detecting phonetic information distinguishing English categories from similar Mandarin categories.

The results of the training study indicated that the English vowel identification training I used caused learners from both groups to improve in their ability to identify English vowel contrasts. Perhaps the most important finding is that much of the improvement in identification occurred within English vowel categories that are considered least similar to any Mandarin categories. Before training, there was a near ceiling effect in performance on those vowels that were most similar to Mandarin categories. This seems to indicate that learners were able to bootstrap on L1 categories to successfully identify similar English vowels. The effect of training, then, was to help them better recognize categories that were not as similar to existing Mandarin categories.

Although both groups of learners in the training study improved in their ability to perceive and produce English vowel contrasts, the training study was inconclusive in determining the relative benefit of artificially lengthening vowel stimuli, or deliberately selecting natural stimuli on the basis of relative dissimilarity to the learners' L1. Neither the Lengthened Vowel training condition, nor the Select Vowel training condition appeared to have a measurable effect vis-à-vis naturally varying vowel training stimuli.

There are a number of possible explanations for this null result. One is that the two types of stimuli manipulation I employed simply do not have a meaningful effect. For example, lengthening vowels may not provide learners with a better chance of detecting important spectral information; similarly, less Mandarin-like English production tokens, on average, may be no easier to discern as members of a new category than are those productions that are more Mandarin-like. In the case of the Select Vowel training, my selection process may have been flawed. Gross statistical evidence that a particular token is less Mandarin-like may not actually reflect learner experience. For example, it may be that a large portion of the training tokens of some 'similar' Mandarin-English categories that were deemed less Mandarin-like are actually still sufficiently Mandarin-like to afford no real advantage in perception. In contrast, for 'new' categories, it may be that nearly all production tokens of those categories were sufficiently distant from Mandarin that each was equally useful as evidence for L2 category formation. Had I used a smaller subset, for example, those 10% that were least Mandarin-like, an effect may have been detected; however, I was concerned that I maintain some degree of variability in the training stimuli and selecting a smaller subset may have had adverse effects related to lack of variability. Another possibility that I should have considered in designing the Select Vowel training stimulus set is that some tokens that are less Mandarin-like may consequently be closer to a competing English vowel category boundary, leading to potential within-English ambiguity. If this were the case, advantages associated with a token being dissimilar from Mandarin might be offset by within L2 English confusion. In future research, it might be better to allow learners themselves to select tokens that they find easiest to discriminate. This could be achieved by presenting all stimuli in initial training sessions, then isolating the items that were best identified during the first few training sessions and subsequently using those for the remainder of the training sessions. This might provide learners with more positive experience that could then lead to more rapid category development and reinforcement. This proposal can be viewed as a modification of best-exemplar training (cf. Jongman & Wade, 2007; Pisoni & Lively, 1995). In my version, best-examplars are not defined in terms of ideal native speaker productions of the L2, but in relation to each productions' interaction with the learners' L1 system. That is, for L2 speakers, best-exemplars are those tokens that learners are most able to perceive as being distinct from L1 categories, while not being confusable with other L2 categories. It may be those tokens that initially form the basis for 'new' L2 category. Further research in this area is needed.

A second possibility concerning the null effect of training condition is that although the Select Vowel stimuli were easier to discern, I failed to detect this. It is possible that the training task design was so effective in orienting learners' attention to the vowels being learned that small differences in training conditions became irrelevant. That is, under the training conditions used in this study, all contrasts were equally discerned, despite qualitative differences in the stimuli used. Perhaps using a different task would lead to differences in performance relative to the training condition. For example, following Guion and Pederson (2007a, 2007b) learners' attention could be directed away from the target vowel toward the consonant to see if the lengthened vowel stimuli had the effect of drawing attention back to the vowel.

A final possible explanation for why there was no effect of training condition relates to the length of the training period. It may be that because of the magnitude of training, initial difficulties learners had with particular stimuli may have been overcome. In their study demonstrating the positive effect of orienting attention in a phonetic training task, Guion and Pederson (2007a) state that they deliberately chose to use a short training session. They indicated that a longer period of training might have changed the magnitude of the effect of orienting attention. It is possible that differences between training conditions in my study may have been detectable in the early stages, but that with more training, all participants began detecting differences, regardless of stimulus type. If this were the case, initial differences may have been washed-out by the time of the first identification test.

7.2.2. Transfer of training to new phonetic contexts

Another goal of my study was to determine if training in one CV context would extend to other CV contexts varying by place and manner of articulation. The results of the training study indicated that training learners to identify English vowels in /b, pV/ contexts resulted in improved identification of the English vowels not only in the /b, pV/ context, but also in /g,kV/ and /z, sV/ contexts. However, improvement in identification scores was greatest for /b, pV/ contexts. In addition, there seemed to be some indication that improvement was greater for /g ,kV/ contexts than for /z, sV/ contexts. This suggests that transfer of learning may be easiest in contexts varying by place, and more difficult for contexts varying by manner. Conversely, it may be that because English velars are more similar to Mandarin velars, while the English fricatives are less like Mandarin fricatives, that L1 played a role. In either case, these findings appear to contradict strong claims regarding the position-sensitive nature of L2 speech learning (cf. Flege, 1995). It appears that learning in one context does sometimes transfer to new contexts, confirming findings of studies by Rochet (1995) and Broersma (2005). Additionally, although Mandarin /o/ does not occur in a post-labial context, it appears that this does not limit Mandarin speakers' ability to bootstrap on this Mandarin category in post-labial English contexts.

To be fair, the results of my study do suggest a general effect of position-sensitivity. The differences in the learners' mean performance in each consonantal context suggest that while learning can transfer from one context to another quite quickly for some learners, other learners may require more experience with these vowel contrasts in the training context, the new context, or both before they are able to successfully transfer learning.

7.2.3. Transfer of training to novel speaker's voice

The improvement in the learners' ability to identify English vowels produced by the twenty voices used in the training stimuli extended to novel productions of those vowels by a speaker whose voice was also used during training. More importantly, the effect of training transferred to vowels produced in multiple CV contexts by an English speaker whose voice was not used in training. This ability to generalize to a new voice supports previous findings, such as those reported by Pisoni and Lively (1995), who argued that high variability training leads to the ability to identify contrasts produced by new voices, while training on a single voice may not.

7.2.4. Transfer of vowel identification training to production

Although previous research has sometimes indicated that L2 vowel identification training does not quickly transfer to production (e.g., Wang, 2002), this was not the case in my study. The results of the training study reported in Chapter 4 indicated a significant improvement in production over a three-week training period. As with identification, what is especially notable is that much of the learning occurred in the production of vowels that were 'new' for Mandarin learners. The fact that the learners in the training study rapidly improved in their ability to identify most of the less similar or 'new' English vowels is in striking contrast to the results reported in Chapter 6, which examined production data from an earlier study of naturalistic English vowel learning (Munro et al., 2004; Munro & Derwing, 2007). In that context, most learning was accounted for by improvement in the Mandarin L1 learners' ability to produce English vowel categories most similar to existing Mandarin L1 categories. The fact that learners were not initially as successful in producing those similar English categories may indicate an effect of Mandarin phonotactics; the English CVC stimuli in that study violated Mandarin phonological constraints because of the presence of a coda stop consonant. Consequently, improvement in producing the English vowels may not have been the result of English vowel learning, but rather, the result of learning to perceive and/or produce Mandarin vowel categories in new English contexts that are unfamiliar for Mandarin L1 speakers. Training, then, seems capable of promoting learning of relatively new L2 categories that appear not to be easily learned in the absence of instruction.

7.3. The effect of the L1 on L2 speech learning

Clearly, Mandarin vowel categories have a substantial effect on the ability of Mandarin learners' of English to identify and produce English vowels. My study has provided further support for concepts related to Best's (1995) PAM and Flege's (1995) SLM regarding the nature of L1/L2 phonological interaction. Mandarin vowel categories that were highly similar to a particular English category appeared to transfer almost immediately to the L2. When a Mandarin category was similar to more than one English category, its substitution for two or more English categories caused confusion within the English L2. Categories that were least similar to Mandarin were more difficult to learn; however, with the exception of English /u/, confusion between such 'new' categories and Mandarin L1 categories rarely occurred.

Some interesting learner behavior emerged with regard to vowel duration that was not clearly predicted on the basis of Mandarin or English. In Mandarin, duration is not a clear cue to vowel identification. In English, it can be a cue, at least when prosodic conditions are held constant. In other words, given similar contexts, vowel duration may help to disambiguate some productions of similar categories in English. However, without context providing information about the relative duration of a given vowel in a given word, duration may not be a particularly useful cue. In the context of my study, the syllables used as stimuli were all presented in the sentence frame, "The next word is ___" for the production task, and extracted from the sentence frame for the identification task. Given that the prosodic context was held constant, vowel duration may have provided a cue for vowel identification. In the L2 production data, there was some indication that for some English contrasts (i.e., /ɒ/-/ʌ/ and /u/-/ʊ/) learners were more successful in learning durational differences than spectral differences. These findings are relevant to claims made by Bohn (1995) concerning what he termed the "desensitization hypothesis." In a study of English learners from German, Spanish, and Mandarin backgrounds, Bohn (1995) found that learners often relied on duration as a primary cue to vowel identification for some L2 English contrasts. For example, relative duration differences between members of English /i/-/ɪ/ and /ɛ/-/æ/ pairs can often provide learners with information that can be used for successful vowel identification. Bohn (1995) argues that reliance on duration occurs when L2 learners have been desensitized to spectral differences in L2 contrasts because those differences are not important for L1 category identification. Consequently, when the distribution of two L2 categories is subsumed by the distribution of a single L1 category covering nearly the same perceptual space, confusion between the two L2 categories ensues. Bohn further claims that the tendency for L2 learners to rely on duration in such L2 contexts is a universal principle, regardless of the learner's L1. While the learners in my study appear to make use of duration, it is not at all certain that they rely on it to the exclusion of spectral properties, at least not for both English /u/-/ʊ/ and /ɒ/-/ʌ/ contrasts. In production, even when productions of /u/ were recognized by the statistical model as being more /ʊ/-like, those productions' raw spectral properties indicated that a distinction between /u/ and /ʊ/ was in fact being made; the problem was that the spectral distinction, although in the right direction, was unlike that of native English speakers. The category boundary between /u/ and /ʊ/, for many learners, appeared to fall within the native English speaker /ʊ/ category. Consequently, although the learners were making a spectral distinction, it may not have been evident to a native English speaker, or at least to the English statistical pattern recognition model.

English /ɒ/-/ʌ/ productions by the learners in my study did suggest greater use or learning of duration and less separation in the spectral domain for those learners who had not acquired the contrast. However, the fact that some learners were successful in acquiring the contrast in production suggests that if desensitization exists, it is not insurmountable.

Finally, for the English /i/-/ɪ/ contrast, which Bohn (1995) indicated was especially problematic for Mandarin learners of English, duration did not appear to play a large role in production errors for the learners in my study. Although /ɪ/ was sometimes identified as /i/ on the identification tests, suggesting some perceptual similarity between members of this pair for some Mandarin learners, this was a relatively rare occurrence. In production, errors in /ɪ/ were almost never recognized as /i/. It was far more common that productions of /ɪ/ were incorrectly recognized as /ɛ/, or even /e/. This suggests that in the CV context, spectral properties were relatively easy to discern for the /i/-/ɪ/ contrast. Overall, the results of my training study suggest that learners were sensitive to many spectral differences, despite some English vowel categories reportedly being subsumed by a single L1 category.

An alternative explanation for differences found in the use of duration vis-à-vis spectral properties in L2 English phonological learning may be that spectral differences are inherently more difficult to learn than are duration distinctions, regardless of interactions with L1 categories. In such an account, rather than 'relying' on duration differences as Bohn (1995) claims, learners may have simply learned duration differences first because they are easier to learn. Claims regarding learner reliance on duration to the exclusion of spectral properties need to be tested in research that holds duration constant for such contrasts. Wang and Munro (2004) manipulated duration in synthetic speech training stimuli in such a way that duration no longer served as an unambiguous cue to vowel identification. For example, some synthetic tokens of English /i/ were deliberately short, while some synthetic tokens of English /ɪ/ were quite long. Wang and Munro (2004) found that this type of manipulation of training stimuli had a positive effect on the learners' ability to acquire spectral properties of such contrasts. This result suggests that learners are sensitive to spectral properties; they only need to have their attention adequately oriented to this dimension.

Perhaps notions of category goodness such as those posited by Best (1995), Best et al. (2001) and Guion et al. (2000) account for learners' ability to maintain a degree of sensitivity to spectral differences in some contexts, such as English /i/-/ɪ/ and /u/-/ʊ/. In other contexts, such as English /ɒ/-/ʌ/, differences in goodness of fit to a single L1 category may be smaller, and therefore the ability to discern these differences is weaker.

7.4. Implications for L2 speech learning models

For the most part, my dissertation research has supported the basic tenets of current L2 speech learning models. Where the results of this research make the largest contribution is in the measurement of the phonetic valence of individual tokens that comprise phonetic categories. While I have used the term 'category' throughout this dissertation as a means of framing my research in historical as well as practical terms, the results clearly point to a need to move away from viewing the beginning of L2 phonological learning in terms of native speaker monolithic 'categories.' This shift has long been present in theory – both the SLM and PAM recognize that not all productions of the same category are equally good members of L2 categories. However, in practice, this theoretical claim is rarely applied. The results of the current study suggest that the development of L2 phonological categories is achieved on the basis of learners assigning individual production tokens to either a similar L1 category, or to an emerging L2 category that is formed on the basis of tokens that do not fit well into any L1 category. In the initial stages of L2 learning, some tokens produced by native speakers of the L2 may be labeled by the learner as 'potentially belonging to a new L2 category' but can be used to establish 'new' L2 categories only after sufficient evidence emerges that they belong to a unique class of sounds. If this reflects reality, such organization of phonological experience in the L2 is not unlike the process infants appear to apply in developing categories in their L1. However, given the fact that L2 learners need to organize L2 phonological experience in the context of an already existing L1 system, it is not surprising that the task is much more complicated and takes much longer than the six months to a year that laying a solid foundation for L1 phonological development typically requires (Polka & Werker, 1994).

Using terms such as 'equivalence classification' (Flege, 1995) and 'functional equivalence classes' (Best et al., 2001), SLM and PAM both treat individual L2 tokens as varying in the degree to which they fit L1 phonological classes or categories. My Metamodel makes the interaction of individual tokens with L1 categories more explicit. The most striking finding from the identification training and production test results, reported in Chapter 4, was the extent to which the learners' responses to each of the two stimulus voices patterned together across learners. Similarities in learners' identification and production responses to the same stimulus provide clear evidence that responses to particular voices vary in predictable ways across a group of learners from the same L1; they respond differently to different tokens of the same L2 category produced by different speakers.

Although *all* learners did not respond in exactly the same way to each stimulus, patterns were evident that suggest a direct interaction between variation in native speaker English productions of a single vowel, and more stable Mandarin speaker L1 vowel categories. Clearly, for some English vowels, one stimulus voice was better, while for other vowels, the other stimulus voice was better. This suggests that learners might benefit more from one voice than another in their development of specific L2 phonological categories. Furthermore, specific production tokens that are not well-identified by the L2 learners may actually hinder development of a particular category. Only after L2 categories have begun to strengthen and become relatively stable can learners begin to benefit from incorporating more variable speech productions. This view is also supported by Jongman and Wade (2007) who indicated that for some L2 categories, initially learning from prototypical examples may better facilitate initial learning. A final ramification of this finding is that the use of an auditory prompt for elicited imitation tasks can result in incorrect conclusions concerning an L2 learner's phonological system. Some learners may respond correctly to an auditory prompt produced by one voice, but incorrectly on a prompt produced by another voice.

7.5. Pedagogical implications

Although this study failed to further our understanding of the effect of manipulating stimuli in phonetic training, it does provide pedagogically relevant information. It appears that given the right sort of task, where attention is explicitly oriented toward phonetic contrasts, learning will occur. This is a positive finding, suggesting that the ability to successfully train some L2 learners to better discern and produce new phonetic contrasts does not require a pedagogically complex or impractical design, although high variability in the tokens and voices used may be necessary.

Another pedagogical implication is that for some highly 'similar' vowels, substantial training is not necessarily important. It appears that learners are able to bootstrap on some L1 categories and substitute these categories for the 'similar' L2 category without any detrimental effect on that production's intelligibility as the intended L2 category. However, while positive transfer of L1 categories to the L2 occurs, it may still be beneficial to include in the training program some exposure to those L2 categories that are nearly identical to L1 categories. It is possible that positive reinforcement of nearly identical or similar categories may actually serve to remind learners of differences between those categories and unfamiliar L2 categories. For example, while Mandarin /a/ might usually suffice as a substitute for English /ɒ/, we cannot conclude that learners do not need training on English /ɒ/, since they need to have it reinforced vis-à-vis English /ʌ/ and /æ/ with which it is sometimes confused. However, for a category such as English /i/, because it was rarely if ever confused with any other category, we can assume that training may be less necessary and that for practical purposes, Mandarin /i/ will permanently suffice in place of English /i/.

Finally, the benefit of instruction should always be considered in terms of its ability to increase the intelligibility of the speaker. Instruction should not be motivated by a desire to rid the speaker of a detectable accent. While my research indicates that the ability to detect spectral differences in L2 vowels appears to remain intact for some adult learners, L2 learners' improvement toward ultimately nativelike perception and production is still a time-consuming task likely requiring years of experience. Consequently, it would be inappropriate to allot scarce instructional time beyond what is necessary to improve the learners' overall intelligibility. Previous research has demonstrated that only a moderate correlation exists between accent and intelligibility (Derwing & Munro, 1997; Munro & Derwing, 1999). An L2 speaker may have a strong accent, yet be highly intelligible. Since the ultimate goal of SLA is successful communication, for most learners, intelligibility should be the final goal. Again, this may indicate that priority in vowel identification and production training should be given to contrasts that result in a lack of intelligibility. Even within the L2, there may be some contrasts that are not particularly important. Brown (1995) argues that all possible minimal pairs within a language do not carry the same communicative weight (i.e., they have a different functional load). Munro and Derwing (2006) have found evidence that L2 phonological errors with a high functional load affect both the perception of accentedness as well as the comprehensibility of an utterance, while phonological errors with a smaller functional load appear to have little impact. The Mandarin learners in my study had particular difficulty with the /u/-/ʊ/ contrast, a contrast that has a relatively small functional load; there are very few instances in English where /u/-/ʊ/ contrast lexically. Therefore, if Mandarin learners go on producing something that is recognized as more /ʊ/-like, when they intend it to be an English /u/, the possibility of such an error resulting in a communicative breakdown is relatively remote. The fact that so few /u/-/ʊ/ lexical contrasts exist in English may also explain why it is so difficult for Mandarin learners of English to develop this contrast. Apart from a deliberately controlled training environment using nonce words such as I employed in the current study, L2 English learners receive virtually no positive or negative evidence that /u/ and /ʊ/ are not simply allophones of the same English category. If this is true, it would be reasonable for Mandarin learners of English to substitute Mandarin /ɤ/ for English /ʊ/ to which it is similar, and by extension, to English /u/, to which it is less similar.

7.6. Further research

The degree to which the Metamodel truly reflects key aspects of the interlanguage of the L2 learner is uncertain. However, the number of accurate predictions this model makes in terms of real-world L2 learner behaviour suggests that it is quite successful in measuring Mandarin-English vowel similarity. Further research examining different language groups is needed to determine whether results will be equally robust in the context of different L1/L2 pairs. If the model is found to be reliable in the context of other language pairings, its validity will be indirectly strengthened.

Additionally, as Hillenbrand and Nearey (1999) point out, successful classification of speech sounds on the basis of spectral information does not prove that human listeners rely on the same information in making categorial identification judgments. Hence, they suggest that pattern recognition studies of the type described in this dissertation must be followed by perceptual experiments comparing the results with judgments of human listeners. While this cannot be directly investigated for the Metamodel because of its idealized and therefore theoretical status, the recognition of the Mandarin accented L2 productions by the English Model can and should be tested against human listeners' responses to the same data. We know that statistical pattern recognition models are effective in classifying L1 data. However, the results of classifying L2 accented speech using the English Model needs to be compared to responses by human listeners to the same production data. Munro and Derwing's (2007) evaluation of the same data using four human listeners provides a preliminary indication that native speakers do not easily agree on what category many accented productions fall into. Their results suggest that there may be a greater likelihood that accented productions will be ambiguous to native speakers than is the case with unaccented productions (c.f. Nearey & Assmann, 1986).

Finally, as mentioned in the preceding discussion, research is needed to determine if training L2 speakers on production tokens that are easiest to identify will help them to establish 'new' L2 categories more quickly. This approach may be able to test my original hypothesis that productions that are less likely to assimilate to the learners' L1 category will promote faster learning.

Bibliography

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., & Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: the case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics, 32,* 233-250.

Andruski, J. E., & Nearey, T. M. (1992). On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables. *Journal of the Acoustical Society of America, 91,* 390-410.

Assmann, P., & Katz, W. (2000). Time-varying spectral change in the vowels of children and adults. *Journal of the Acoustical Society of America, 108,* 1856-1866.

Assmann, P., Nearey, T. M., & Hogan, J. (1982). Vowel identification: orthographic, perceptual, and acoustic aspects. *Journal of the Acoustical Society of America, 71,* 975-989.

Beddor, P. S., & Gottfried, T. L. (1995). Methodological issues in cross-linguistic speech perception research with adults. In W. Strange (Ed.), *Speech perception in linguistic experience: Theoretical and methodological issues* (pp. 207-232). Timonium, MD: York Press.

Best, C. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience.* Timonium, MD: York Press, pp. 171-204.

Best, C. T., McRoberts, G. W., & Goodwell, E. (2001). Discrimination of non-native contrasts varying in perceptual assimilation to the listener's native phonological system. *Journal of the Acoustical Society of America, 109,* 775-794.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1989). The phonological basis of perceptual loss for non-native contrasts: maintenance of discrimination among Zulu clicks by English-speaking adults and infants. *Human Perception and Performance, 14,* 345-360.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementaries. In O.-S. Bohn & M. J. Munro (Eds.), *Second-language speech learning: The role of language experience in speech perception and production: A festschrift in honour of James E. Flege* (pp. 13-34). Amsterdam: John Benjamins.

Birdsong, D. (1992). Ultimate attainment in second language acquisition. *Language, 68,* 706-755.

Birdsong, D. (2007). Nativelike pronunciation among late learners of French as a second language. In O.-S. Bohn & M. J. Munro (Eds.), *Second-language speech learning: The role of language experience in speech perception and production: A festschrift in honour of James E. Flege* (pp. 99-116). Amsterdam: John Benjamins.

Broersma, M. (2005). Perception of familiar contrasts in unfamiliar positions. *Journal of the Acoustical Society of America, 117,* 3890-3901.

Brown, A. (1995). Minimal pairs: minimal importance? *ELT Journal*, 49, 169-175.

Bohn, O.-S. (1995). Cross-language speech perception in adults: First language transfer doesn't tell it all. In W. Strange (Ed.), *Speech perception in linguistic experience: Theoretical and methodological issues* (pp. 273-304). Timonium, MD: York Press.

Bohn, O.-S., & Flege, J. E. (1990). Interlingual identification and the role of foreign language experience in L2 vowel perception. *Applied Psycholinguistics, 11,* 303-328.

Bohn, O.-S., & Flege, J. E. (1992). The production of new and similar vowels by adult German learners of English. *Studies in Second Language Acquisition, 14,* 131-158.

Bohn, O.-S., & Polka, L. (2001). Target spectral, dynamic spectral, and duration cues in infant perception of German vowels. *Journal of the Acoustical Society of America, 110,* 504-515.

Bond, Z. S. (1999). *Slips of the Ear. Errors in the perception of casual conversation.* New York: Elsevier.

Bongaerts T., Mennen, S., & van der Slik, F. (2000). Authenticity of pronunciation in naturalistic second language acquisition: The case of very advanced late learners of Dutch as a second language *Studia Linguistica, 54*, 298-308.

Bongaerts, T., van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition, 19,* 447-465.

Borden, G., Gerber, A., & Milsark, G. (1983). Production and perception of the /r/-/l/ contrast in Korean adults learning English. *Language Learning, 33,* 499-526.

Bradlow, R. R., Pisoni, D. B., Akahana-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America, 101,* 2299-2310.

Bybee, J. (2001). *Phonology and language use.* Cambridge, Cambridge University Press.

Bybee, J. (2002). Phonological evidence for exemplar storage of multiword sequences. *Studies in Second Language Acquisition, 24,* 215-221.

Cenoz, J., & Garcia Lecumberri, M. L. (1999). The effect of training on discrimination of English vowels. *International Review of Applied Linguistics, 37*, 261-275.

Chen, M. Y. (1976). From middle Chinese to modern Peking. *Journal of Chinese Linguistics, 4*(2/3). 113-277.

Chen, Y., Robb, M., Gilbert, H., & Lerman, J. (2001). Vowel production by Mandarin speakers of English. *Clinical Linguistics and Phonetics, 15,* 427-440.

Corder, S. P. (1971). Idiosyncratic errors and error analysis. *International Review of Applied Linguistics, 9,* 147-159.

Cross, T. G. (1977). Mothers' speech adjustments: The contributions of selected child listener variables. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to children: Language input and acquisition* (pp. 151-188). Cambridge: Cambridge University Press.

Cross, T. G. (1978). Mothers' speech and its association with rate of linguistic development in young children. In N. Waterson & C. Snow (Eds.), *The development of communication* (pp. 199-216). New York: Wiley.

de Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *ARLO, 4,* 129-134.

Derwing, T. M., Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 20,* 1-16.

Derwing, T. M., Munro, M. J., & Wiebe, G. E. (1997). Pronunciation instruction for "fossilized" learners: Can it help? *Applied Language Learning, 8*, 185-203.

Derwing, T. M., Munro, M. J., & Wiebe, G. E. (1998). Evidence in favour of a broad framework for pronunciation instruction. *Language Learning, 48,* 393-410.

Doughty, C., & Long, M. H. (Eds.). (2003). *Handbook of second language acquisition.* Oxford: Blackwell.

Duanmu, S. (2003). *The Phonology of Standard Chinese.* Oxford: Oxford University Press.

Eckman, F. R. (1977). Markedness and the contrastive analysis hypothesis. *Language Learning, (27),* 315-300.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition, 24*, 143-188.

Ellis, R. (1990). *Instructed second language learning*. Oxford: Blackwell.

Ellis, R. (1994). *The study of second language acquisition.* Oxford: Oxford University Press.

Fernald, A., & Morikawa, H. (1993). Common themes and cultural variations in Japanese and American mothers' speech to infants. *Child Development, 64,* 637-656.

Flege, J. E. (1981). The phonological basis of foreign accent: A hypothesis. *TESOL Quarterly (15),* 443-455.

Flege, J. E. (1984). The detection of French accent by American listeners. *Journal of the Acoustical Society of America, (76),* 692-707.

Flege, J. E. (1987). Equivalence classifications of L1 and L2 sounds in second language acquisition. In A. James & J. Leather (Eds.), *Sound patterns in second language acquisition.* Dordrecht, Foris, 9-39.

Flege, J. E. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception in linguistic experience: Theoretical and methodological issues* (pp. 229-273). Timonium, MD: York Press.

Flege, J. E. (2005, May 14). Origins and development of the Speech Learning Model. Paper presented at the 1st ASA Workshop on L2 Speech Learning. Vancouver, BC: Simon Fraser University.

Flege, J. E., Bohn, O-S. (1989). The perception of English vowels by native speakers of Spanish. *Journal of the Acoustical Society of America, 85,* Supplement 1, S85.

Flege, J. E., Bohn, O-S., & Jang, S. (1997). Effeets of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics, 25,* 437-470.

Flege, J. E., Frieda, E. M., & Nozawa, T. (1997). Amount of native-language (L1) use affects the pronunciation of an L2. *Journal of Phonetics, 25,* 169-186.

Flege, J. E. and Hillandbrand, J. M. (1984). Limits on phonetic accuracy in foreign language speech perception. *Journal of the Acoustical Society of America, 76,* 708-721.

Flege, J. E., MacKay, I. R. A., & Meador, D. (1999). Native Italian speakers' production and perception of English vowels. *Journal of the Acoustical Society of America, 106*, 2973-2987.

Flege, J. E., & Munro, M. J. (1994). The word unit in L2 speech production and perception. *Studies in Second Language Acquisition, 16,* 381-411.

Flege, J. E., Munro, M. J., & Fox, R. (1994). Auditory and categorical effects on cross-language vowel perception. *Journal of the Acoustical Society of America, 95,* 3623-3641.

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America, 97,* 3125-3134.

Flege, J. E., Schirru, C., & MacKay, I. R. A. (2003). Interaction between the native and second language phonetic subsystems. *Speech Communication, 40,* 467-491.

Fowler, C. A., (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14, 3-28.

Gass, S. (1984). Development of speech perception and speech production abilities in adult second language learners. *Applied Psycholinguistics, 5*, 51-74.

Gass, S. M., & Alvarez, M. J. (2005). Attention when? *Studies in Second Language Acquisition, 27*, 1-31.

Gerken, L., & Aslin, R. N. (2005). Thirty years of research on infant speech perception: the legacy of Peter W. Jusczyk. *Language Learning and Development, 1,* 5-21.

Goldinger, S. D. (1997). Words and voices: Perception and production in an episodic lexicon. In K. Johnson & J. Mullennex (Eds.), *Talker variability in speech processing* (pp. 33-66). New York: Academic Press.

Gottfried, T. L., & Suiter, T. L. (1997). Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones. *Journal of Phonetics, 25,* 207-231.

Guenther, F. H. (2000). An analytical error invalidates the "depolarization" of the perceptual magnet effect. *Journal of the Acoustical Society of America, 107,* 3576-3580.

Guion, S. G., Flege, J. E., Akahane-Yamada, R., & Pruitt, J. C. (2000). An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *Journal of the Acoustical Society of America, 107,* 2711-2724.

Guion, S. G., & Lee, B. (2006). The role of phonetic processing in second language acquisition. *English Language and Linguistics, 21,* 123-148.

Guion, S. G., & Pederson, E. (2002). The role of orienting attention for learning novel phonetic categories. *The Institute of Cognitive and Decision Sciences at the University of Oregon, Technical Reports,* No. 02-6.

Guion, S. G., & Pederson, E. (2007a). Investigating the role of attention in phonetic learning. In O.-S. Bohn & M. J. Munro (Eds.), S*econd-language speech learning: The role of language experience in speech perception and production: A festschrift in honour of James E. Flege* (pp. 57-77). Amsterdam: John Benjamins.

Guion, S. G., & Pederson, E. (2007b, under review). Orienting attention during phonetic training facilitates learning.

Guion, S. G., & Pederson, E. (unpublished ms). The effect of attention on training non-native listeners to discriminate Hindi segments: A first study. *University of Oregon: Eugene.*

Hardison, D. (2003). Acquisition of second-language speech: effects of visual cues, context, and talker variability. *Applied Psycholinguistics, 24,* 495-522.

Hillenbrand, J., Getty, L., Clark, M. & Wheeler, L., (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America, (97),* 3099-3111.

Hillenbrand, J. M., & Nearey, T. M. (1999). Identification of resynthesized /hVd/ utterances: Effects of formant contour. *Journal of the Acoustical Society of America, 105,* 3509-3523.

Liu, H. -M., Kuhl, P. K., & Tsao, F. -M. (2002, December). Mother's exaggerated acoustic-phonetic characteristics in infant-directed speech are highly correlated with infant's speech discrimination skills in the first year of life. Poster presented at the *144th Meeting of the Acoustical Society of America, First Pan-American/Iberian Meeting on Acoustics*, Cancun, Mexico.

Ingram, J., & Park, S. G. (1997). Cross-language vowel perception and production by Japanese and Korean learners of English. *Journal of Phonetics, 25,* 343-370.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2002). A perceptual interference account of acquisition difficulties of non-native phonemes. *Cognition, 87,* B47-B57.

Jamieson, D. G., & Morosan, D. E. (1986). Training non-native speech contrasts in adults: Acquisition of the English /δ/ - /θ/ contrast by francophones. *Perception and Psychophysics, 40,* 205-215.

Jamieson, D. G., & Morosan, D. E. (1989). Training new, nonnative speech contrasts: A comparison of the prototype and perceptual fading techniques, *Canadian Journal of Psychology, 43,* 88-96.

Jenkins, J. J., Strange, W., & Trent, S. A. (1999). Context-independent dynamic information for the perception of coarticulated vowels. *Journal of the Acoustical Society of America, 106,* 438 - 448.

Jia, G., Strange, W., Wu, Y., Collado, J., & Guan, Q. (2006). Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure. *Journal of the Acoustical Society of America, 119*, 1118-1130.

Jongman, A., & Wade, T. (2007). Acoustic variability and perceptual learning: The case of non-native accented speech. In O.-S. Bohn & M. J. Munro (Eds.), S*econd-language speech learning: The role of language experience in speech perception and production: A festschrift in honour of James E. Flege* (pp. 133-166). Amsterdam: John Benjamins.

Jusczyk, P. W. (1997). *The discovery of spoken language.* Cambridge, MA: MIT Press.

Kewley-Port, D., Watson, C. S., & Foyle, D. C. (1988). Auditory temporal acuity in relation to category boundaries; speech and nonspeech stimuli. *Journal of the Acoustical Society of America, 83,* 1133-1145.

Kingston, J., & Diehl, R. (1995). Phonetic knowledge. *Language, 70,* 419-454.

Kirchhoff, K., & Schimmel, S. (2005). Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. *Journal of the Acoustical Society of America, 117,* 2238-2246.

Kuhl, P. K., & Iverson, P. (1995). Linguistic experience and the "Perceptual magnet effect". In W. Strange (Ed.), *Speech perception in linguistic experience: Theoretical and methodological issues* (pp. 121-154). Timonium, MD: York Press.

Kuhl, P. K. (2004). Cracking the speech code. *Neuroscience, 5,* 831-843.

Lado, R. (1957*). Linguistics across cultures.* Ann Arbor, University of Michigan Press.

Lenneberg, E. (1967). *Biological foundations of language*. New York: Wiley.

Leow, R. P. (1997). Attention, awareness, and foreign language behaviour. *Language Learning, 47,* 467-505.

Lee, J. G., Cadierno, T., Glass, W. R., & VanPatten, B. (1997) The effects of lexical and grammatical cues on processing past temporal reference in second language input. *Applied Language Learning, 8*, 1-23.

Lee, W.-S. and Zee, E. (2003). Standard Chinese (Beijing). *Journal of the International Phonetics Association, 33,*(3), 109-112.

Li, C. N., & Thompson, S. A. (1997). *Mandarin Chinese: A functional reference grammar.* Taipei, Taiwan: The Crane Publishing Co.

Liberman, A. M., & Mattingly, I. G. (1985). The Motor Theory of speech perception revised. *Cognition, 21,* 1-36.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, 89, 874-886.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1993). Training listeners to perceive novel phonetic categories: How do we know what is learned? *Journal of the Acoustical Society of America*, 94, 1148-1151.

Logan, J. S., & Pruitt, J. S. (1995). Methodological issues in training listeners to perceive non-native phonemes. In W. Strange (Ed.), S*peech perception in linguistic experience: Theoretical and methodological issues* (pp. 351-377). Timonium, MD: York Press.

Long, M. (1983). Does second language instruction make a difference? A review of the research. *TESOL Quarterly* 17: 359-382.

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition, 19,* 37-66.

Maddieson, I. (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.

Major, R., & Kim, E. (1996). The similarity differential rate hypothesis. *Language Learning, 46,* 465-496.

McAllister, R. (2001). Experience as a factor in L2 phonological acquisition. *Lund University, Department of Linguistics Working Papers, 49,* 116-119.

McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the r-l contrast to Japanese adults: predictions of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, and Behavioral Neuroscience 2, (2),* 89-108.

McClelland, J. L., Fiez, J. A., & McCandliss, B. D., (2002). Teaching the non-native [r]-[l] speech contrast to Japanese adults: Training methods, outcomes, and neural basis. *Physiology and Behavior, 77,* 657-662.

Mochizuki, M. (1981). The identification of /r/ and /l/ in natural and synthesized speech. *Journal of Phonetics, 9,* 283-303.

Morrison, G. S. (2006). L1 and L2 production and perception of English and Spanish vowels: A statistical modeling approach. *Unpublished doctoral dissertation.* Edmonton, Alberta: University of Alberta.

Moyer, A. (1999). Ultimate attainment in L2 phonology: The critical factors of age, motivation and instruction. *Studies in Second Language Acquisition, 21,* 81-108.

Munro, M., & Mann, V. (2005). Age of immersion as a predictor of foreign accent. *Applied Psycholinguistics, 26,* 311-341.

Munro, M. J. (1993). Productions of English vowels by native speakers of Arabic: Acoustic measurements and accentedness ratings. *Language and Speech, (36),* 39-66.

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 49,* (Supplement 1), 285-310.

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System, 34,* 520-531.

Munro, M. J., & Derwing, T. M. (2007, accepted for publication). A developmental study of Mandarin and Russian speakers' English vowels. *Language Learning.*

Munro, M. J., Derwing, T. M., & Thomson, R. I. (2003). A longitudinal examination of English vowel learning by Mandarin speakers. *Canadian Acoustics, 31*, (3), 32-33. Proceedings of the annual conference of the Canadian Acoustics Association. Edmonton, AB.

Munro, M. J., Flege, J. E., & MacKay, I. R. A. (1996). The effect of age of second-language learning on the production of English vowels. *Applied Psycholinguistics, 17,* 313-334.

Murray, A. D., Johnson, J., & Peters, J. (1990). Fine-tuning of utterance length to preverbal infants: Effects on later language development. *Journal of Child Language, 17*, 511-525.

Nearey, T. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America, 101,* 3241-3254.

Nearey, T. M. (2004). On the factorability of phonological units in speech perception. In J. Local, R. Ogden -and R. Temple (Eds.), *Phonetic interpretation: Papers in laboratory phonology VI* (pp. 197-221). Cambridge: Cambridge University Press.

Nearey, T. M., & Assmann, P. (1986). Modeling the role of vowel inherent spectral change in vowel identification. *Journal of the Acoustical Society of America, 80,* 1297- 308.

Ochs, E., & Schieffelin, B. (1994). Language acquisition and socialization: Three developmental stories and their implications. In B. Blount (Ed.), *Language, culture, and society: A book of readings* (pp. 470-512). Long Grove, IL: Waveland Press.

Patkowski, M. (1990). Age and accent in a second language: A reply to James Emil Flege, *Applied Linguistics, 11,* 73-89.

Pica, T. (1985). The selective impact of classroom instruction on second-language acquisition. *Applied Linguistics, 6,* 214-222.

Pierrehumbert, J. B. (2001). Exemplar-dynamics. Word frequency, lenition and contrast. In J. Bybee, & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 137-157). Philadelphia: John Benjamins.

Piske, T., MacKay, I. R. A., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics, 29,* 191-215.

Pisoni, D. B., & Lively, S. E. (1995). Variability and invariance in speech perception: A look at some old problems in perceptual learning. In W. Strange (Ed.), *Speech perception in linguistic experience: Theoretical and methodological issues* (pp. 433-459). Timonium, MD: York Press.

Polka, L. (1995). Linguistic influences in adult perception of non-native vowel contrasts. *Journal of the Acoustical Society of America, 97,* 1286-1296.

Polka, L., & Bohn, O.-S. (1996). A cross-language comparison of vowel perception in English-learning and German-learning infants. *Journal of the Acoustical Society of America, 100,* 577-592.

Polka, L., & Werker, J. F. (1994). Developmental changes in the perception of non-native vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 421-435.

Posner, M. I., & Peterson, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience, 13,* 25-42.

Rochet, B. L. (1995). Perception and production of second-language speech sounds by adults. In W. Strange (Ed.), *Speech perception in linguistic experience: Theoretical and methodological issues* (pp. 379-410). Timonium, MD: York Press.

Robinson, P. (1995). Review article: Attention, memory, and the "noticing" hypothesis. *Language Learning, 45,* 283-331.

Robinson, P. (2003). Attention and memory during SLA. In C. Doughty & M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 631-678). Oxford: Blackwell.

Scovel, T. (1988). *A time to speak. A psycholinguistic inquiry into the critical period for human speech.* Rowley, MA: Newbury House.

Scovel, T. (1969). Foreign accents, language acquisition, and cerebral dominance. *Language Learning, 19,* 245-253.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11,* 129-158.

Schmidt, R. (1993). Awareness and second language acquisition. *Annual Review of Applied Linguistics, 13,* 206-226.

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 1-63). Honolulu: University of Hawaii Press.

Schmidt, R. (2001). Attention. In P. Robinson. (Ed.), *Cognition and second language instruction.* (pp. 3-32). Cambridge: Cambridge University Press.

Schumann, J. (1975). Affective factors and the problem of age in second language acquisition. *Language Learning, 25,* 209-225.

Schumann, J. (1978). *The pidginization process: A model for second language acquisition.* Rowley, MA: Newbury House.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics, 10,* 209-231.

Sharwood Smith, M. (1993). Input enhancement in instructed SLA: Theoretical bases. *Studies in Second Language Acquisition,* 15, 165-179.

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistcs, 3,* 243-261.

Simard, D., & Wong, W. (2001). Alertness, orientation, and detection: The conceptualization of attentional functions in SLA. *Studies in Second Language Acquisition, 23,* 103-124.

Smith, G. P. (2000). *Canadian language benchmarks, 2000.* Ottawa, ON: Citizenship and Immigration Canada.

Strange, W. (1989). Evolving theories of vowel perception. *Journal of the Acoustical Society of America, 85,* 2081-2087.

Strange, W. (1995). Cross-language studies of speech perception. A historical review. In W. Strange (Ed.), S*peech perception in linguistic experience: Theoretical and methodological issues* (pp. 3-45).

Strange, W. (2007). Cross-language similarity of vowels: Theoretical and methodological issues. In O.-S. Bohn & M. J. Munro (Eds.), S*econd-language speech learning: The role of language experience in speech perception and production: A festschrift in honour of James E. Flege* (pp. 35-55). Amsterdam: John Benjamins.

Strange, W., Bohn, O.-S., Trent, S.A., & Nishi, K. (2004). Acoustic and perceptual similarity of North German and American English vowels. *Journal of the Acoustical Society of America, 115,* 1791-1807.

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception and Psychophysics, 36,* 131-145. Timonium, MD: York Press.

Thompson, I. (1991). Foreign accents revisited: The English pronunciation of Russian immigrants. *Language Learning, 41,* 177-204.

Thomson, R. I. (2003). L2 vowel perception: Perceptual assimilation to what? *Canadian Acoustics, 31,* (3), 36-37. Proceedings of the annual conference of the Canadian Acoustics Association. Edmonton, AB.

Thomson, R. I. (2005). A pattern recognition approach to English L2 vowel learning. Unpublished Generals Paper II. Edmonton, AB: University of Alberta.

Tomlin, R. S., & Villa, V. (1994). Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition, 16,* 183-203.

VanPatten, B. (1996). *Input processing and grammar instruction.* New York: Ablex.

VanPatten, B. (2002). Processing instruction: An update. *Language Learning, 52,* 755-803.

Wang, X. (2002). *Training Mandarin and Cantonese speakers to identify English vowel contrasts: long term retention and effects on production.* Unpublished PhD. Dissertation. Burnaby, BC: Simon Fraser University.

Wang, X., & Munro, M. J. (2004). Computer-based training for learning English vowel contrasts. *System, 32,* 539-552.

Werker, J. F. (1995). Age-related changes in cross-linguistic speech perception: Standing at the crossroads. In W. Strange (Ed.), S*peech perception in linguistic experience: Theoretical and methodological issues* (pp. 155-169). Timonium, MD: York Press.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development, 1,* 197-234.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7,* 49-63.

Yamada, R. E. (1995). Age and acquisition of second language speech sounds perception of American English /r/ and /l/ by native speakers of Japanese. In W. Strange (Ed.), *Speech perception in linguistic experience: Theoretical and methodological issues* (pp. 305-319). Timonium, MD: York Press.

Zampini, M. L., & Green, K. P. (2001). The voicing contrast in English and Spanish: The relationship between perception and production. In J. L. Nicol. (Ed.), *One mind, two languages: Bilingual language processing.* (pp., 23-48). Oxford: Blackwell.

Zhang, C., & Nearey, T. M. (2003, Oct 16). Acoustic cues to voicing of initial consonants in Mandarin CV syllables. *Paper presented at the annual conference of the Canadian Acoustical Association*, Edmonton, AB.

Appendix 1. Alternate pattern recognition model results: Mandarin CV with vowel duration; English CV without vowel duration; Metamodel CV with vowel duration

Table A1.1. Mandarin Model trained and tested on native speaker Mandarin productions with vowel duration included as a variable.

		Vowel identified by Mandarin pattern recognition model						
		/i/	/e/	/a/	/uə/	/o/	/ɤ/	/u/
Intended Mandarin vowels repeated in response to auditory stimuli	/i/	**100**	--	--	--	--	--	--
	/e/	2.5	**97.5**	--	--	--	--	--
	/a/	--	--	**100**	--	--	--	--
	/uə/	--	--	--	**82.5**	7.5	10	--
	/o/	--	--	--	--	**80**	5	15
	/ɤ/	--	--	--	22.5	--	**77.5**	--
	/u/	--	--	--	--	--	--	**100**
Total Correct		91% (94% without vowel duration)						

Table A1.2. English Model trained and tested on native speaker English productions without vowel duration included as variable.

		Vowel identified by English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended English vowels repeated in response to auditory stimuli	/i/	**95**	--	5	--	--	--	--	--	--	--
	/ɪ/	--	**87.5**	--	12.5	--	--	--	--	--	--
	/e/	--	2.5	**97.5**	--	--	--	--	--	--	--
	/ɛ/	--	12.5	--	**80**	7.5	--	--	--	--	--
	/æ/	--	--	--	10	**85**	2.5	2.5	--	--	--
	/ɒ/	--	--	--	--	2.5	**72.5**	25	--	--	--
	/ʌ/	--	--	--	--	15	12.5	**60**	--	12.5	--
	/o/	--	--	--	--	--	--	--	**97.5**	--	2.5
	/ʊ/	--	--	--	5	--	2.5	2.5	--	**90**	--
	/u/	--	--	--	--	--	--	--	2.5	--	**97.5**
Total correct		86% (91% with vowel duration cue)									

Table A1.3. Metamodel trained and tested on native speaker English and Mandarin productions with vowel duration included as a variable.

		Vowel recognized by Metamodel																
		English										Mandarin						
		/i/$_{e}$	/ɪ/$_{e}$	/e/$_{e}$	/ɛ/$_{e}$	/æ/$_{e}$	/ɒ/$_{e}$	/ʌ/$_{e}$	/o/$_{e}$	/ʊ/$_{e}$	/u/$_{e}$	/i/$_{m}$	/e/$_{m}$	/a/$_{m}$	/uə/$_{m}$	/o/$_{m}$	/ɤ/$_{m}$	/u/$_{m}$
Intended vowels produced in English or Mandarin English	/i/$_{e}$	**60**	--	5	--	--	--	--	--	--	--	35	--	--	--	--	--	--
	/ɪ/$_{e}$	--	**90**	--	10	--	--	--	--	--	--	--	--	--	--	--	--	--
	/e/$_{e}$	--	--	**87.5**	--	--	--	--	--	--	--	--	12.5	--	--	--	--	--
	/ɛ/$_{e}$	--	10	--	**82.5**	7.5	--	--	--	--	--	--	--	--	--	--	--	--
	/æ/$_{e}$	--	--	--	5	**90**	--	2.5	--	--	--	--	--	2.5	--	--	--	--
	/ɒ/$_{e}$	--	--	--	--	--	**72.5**	--	--	--	--	--	--	27.5	--	--	--	--
	/ʌ/$_{e}$	--	--	--	--	2.6	2.6	**74.4**	--	7.7	--	--	--	12.8	--	--	--	--
	/o/$_{e}$	--	--	--	--	--	--	--	**57.5**	--	2.5	--	--	--	--	40	--	--
	/ʊ/$_{e}$	--	--	--	2.5	2.5	--	2.5	--	**80**	--	--	--	--	--	--	12.5	--
	/u/$_{e}$	--	2.5	--	--	--	--	--	--	--	**95**	--	--	--	--	--	--	2.5
Mandarin	/i/$_{m}$	27.5	--	--	--	--	--	--	--	--	--	**72.5**	--	--	--	--	--	--
	/e/$_{m}$	2.5	--	30	--	--	--	--	--	--	--	--	**67.5**	--	--	--	--	--
	/a/$_{m}$	--	--	--	--	2.5	20	12.5	--	--	--	--	--	**65**	--	--	--	--
	/uə/$_{m}$	--	--	--	--	--	--	--	--	--	--	--	--	--	**82.5**	5	10	2.5
	/o/$_{m}$	--	--	--	--	--	--	--	25	2.5	2.5	--	--	--	--	**60**	2.5	7.5
	/ɤ/$_{m}$	--	--	--	5	--	--	--	--	25	--	--	--	--	25	--	**45**	--
	/u/$_{m}$	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	**100**
Total Correct		74.5% (73% with duration cue)																

Table A1.4. English items tested on the Mandarin Model with vowel duration included as a variable.

		Vowel identification by Mandarin Model						
		Percentage of tokens (Average APP)						
		/i/$_{m}$	/e/$_{m}$	/a/$_{m}$	/uə/$_{m}$	/o/$_{m}$	/ɤ/$_{m}$	/u/$_{m}$
Vowel tokens produced by NSs in response to English stimuli	/i/$_{e}$	95 (100)	5 (.92)	--	--	--	--	--
	/ɪ/$_{e}$	--	15 (.90)	--	--	--	85 (.98)	--
	/e/$_{e}$	--	100 (1.00)	--	--	--	--	--
	/ɛ/$_{e}$	--	2.5 (.95)	50 (.89)	--	--	47.5 (.95)	--
	/æ/$_{e}$	--	--	100 (1.00)	--	--	--	--
	/ɒ/$_{e}$	--	--	100 (.99)	--	--	--	--
	/ʌ/$_{e}$	--	--	82.5 (.96)	--	--	17.5 (.98)	--
	/o/$_{e}$	--	--	--	--	100 (.96)	--	--
	/ʊ/$_{e}$	--	--	2.5 (1.00)	--	--	97.5 (.97)	--
	/u/$_{e}$	--	--	--	--	20 (.77)	37.5 (.87)	42.5 (.83)

Table A1.5. Mandarin items tested on the English Model. without vowel duration included as a variable.

		Vowel identification by English Model Percentage of tokens (average APP)									
		/i/$_e$	/ɪ/$_e$	/e/$_e$	/ɛ/$_e$	/æ/$_e$	/ɒ/$_e$	/ʌ/$_e$	/o/$_e$	/ʊ/$_e$	/u/$_e$
Vowel produced in response to Mandarin stimuli	/i/$_m$	100 (1.00)	--	--	--	--	--	--	--	--	--
	/e/$_m$	2.5 (1.00)	--	97.5 (1.00)	--	--	--	--	--	--	--
	/a/$_m$	--	--	--	--	17.5 (.63)	77.5 (.84)	5 (.77)	--	--	--
	/uə/$_m$	--	--	--	--	--	--	--	7.5 (1.00)	92.5 (.97)	--
	/o/$_m$	--	--	--	--	--	--	--	97.5 (.96)	2.5 (.64)	--
	/ɤ/$_m$	--	--	--	5 (.94)	--	--	--	--	95 (.99)	--
	/u/$_m$	--	--	--	--	--	--	--	67.5 (.95)	7.5 (.95)	25 (.89)

Appendix 2. Details of training study participants

Table A2.1. Details of all English vowel training participants who completed the training portion of the study (n = 26).

Long Vowel Group

ID	Gender	Age	Months in Canada	Months of ESL instruction	Yrs of EFL instruction in China
1	F	35	8	3	10
2	M	38	5	3	25
3	F	35	12	12	1
4	F	27	5	3	10
5	F	40	6	1	11
6	F	36	36	3	10
7	M	48	14	13	4
8	M	32	8	4	16
9	M	31	6	3	10
10	F	30	5	3	10
11	M	33	6	3	20
Means		35	10.1	4.6	11.5

Select Vowel Group

ID	Gender	Age	Months in Canada	Months of ESL instruction	Yrs of EFL instruction in China
12	M	37	6	5	10
13	F	31	6	3	8
14	M	40	6	6	7
15	F	32	6	3	10
16	F	50	14	3	1
17	F	41	7	3	4
18	F	37	7	3	10
19	F	40	9	3	6
20	M	36	48	3	13
21	F	35	4	3	10
22	F	37	12	3	10
Means		37.5	12.8	3.9	8.1

Control Vowel Group

ID	Gender	Age	Months in Canada	Months of ESL instruction	Yrs of EFL instruction in China
23	F	33	12	8	6
24	F	38	30	5	10
25	F	29	4	0	8
26	M	38	9	4	7
Means		34.5	13.8	4.3	7.8

Table A2.2. Details of all English vowel training participants who completed the delayed post-test portion of the study (n = 18).

Long Vowel Group

ID	Gender	Age	Months in Canada	Months of ESL instruction	Yrs of EFL instruction in China
1	F	35	8	3	10
2	M	38	5	3	25
3	F	35	12	12	1
5	F	40	6	1	11
6	F	36	36	3	10
7	M	48	14	13	4
8	M	32	8	4	16
10	F	30	5	3	10
11	M	33	6	3	20
Means		36.3	11.1	5.0	11.9

Select Vowel Group

ID	Gender	Age	Months in Canada	Months of ESL instruction	Yrs of EFL instruction in China
13	F	31	6	3	8
14	M	40	6	6	7
15	F	32	6	3	10
16	F	50	14	3	1
17	F	41	7	3	4
18	F	37	7	3	10
19	F	40	9	3	6
20	M	36	48	3	13
21	F	35	4	3	10
Means		38.0	11.9	3.3	7.7

Appendix 3. Details of individual CV production tokens used for Generalization and Production test stimuli

Table A3.1. English Model recognition of Voice 1 /b, pV/ stimuli (vowel duration included). Numbers reflect APPs of each token being a member of the corresponding category. Largest English APPs are in bold.

		APP scores of belonging to each English category									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended vowels produced in English by Stimulus Voice 1	/bi/	**1.00**	--	--	--	--	--	--	--	--	--
	/bɪ/	--	**1.00**	--	--	--	--	--	--	--	--
	/be	--	--	**1.00**	--	--	--	--	--	--	--
	/bɛ/	--	0.14	--	**0.86**	--	--	--	--	--	--
	/bæ/	--	--	--	--	**0.99**	--	--	--	--	--
	/bɒ/	--	--	--	--	--	**0.99**	--	--	--	--
	/bʌ/	--	--	--	--	--	0.01	**0.74**	--	0.25	--
	/bo/	--	--	--	--	--	--	--	**1.00**	--	--
	/bʊ/	--	--	--	--	--	--	--	--	**1.00**	--
	/bu/	--	--	--	--	--	--	--	--	--	**1.00**
	/pi/										
	/pɪ/	**1.00**	--	--	--	--	--	--	--	--	--
	/pe	--	**0.99**	--	0.01	--	--	--	--	--	--
	/pɛ/	--	--	**1.00**	--	--	--	--	--	--	--
	/pæ/	--	--	--	**0.89**	0.01	--	0.10	--	--	--
	/pɒ/	--	--	--	--	**0.98**	0.01	0.01	--	--	--
	/pʌ/	--	--	--	--	--	**0.96**	0.04	--	--	--
	/po/	--	--	--	--	--	0.01	**0.99**	--	--	--
	/pʊ/	--	--	--	--	--	--	--	**1.00**	--	--
	/pu/	--	--	--	--	--	--	--	--	**1.00**	--
		--	--	--	--	--	--	--	--	--	**1.00**

Table A3.2. English Model recognition of Voice 2 /b, pV/ stimuli (vowel duration included). Numbers reflect APPs of each token being a member of the corresponding category. Largest English APPs are in bold.

		APP scores of belonging to each English category									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended vowels produced in English by Stimulus Voice 1	/bi/	**1.00**	--	--	--	--	--	--	--	--	--
	/bɪ/	--	**1.00**	--	--	--	--	--	--	--	--
	/be	--	--	**1.00**	--	--	--	--	--	--	--
	/bɛ/	--	0.05	--	**0.94**	--	--	--	--	--	--
	/bæ/	--	--	--	--	**1.00**	--	--	--	--	--
	/bɒ/	--	--	--	--	--	**1.00**	--	--	--	--
	/bʌ/	--	--	--	--	--	0.11	**0.81**	--	0.08	--
	/bo/	--	--	--	--	--	--	--	**1.00**	--	--
	/bʊ/	--	--	--	--	--	--	--	--	**1.00**	--
	/bu/	--	--	--	--	--	--	--	--	--	**1.00**
	/pi/	**0.97**	--	0.03	--	--	--	--	--	--	--
	/pɪ/	--	**0.99**	--	0.01	--	--	--	--	--	--
	/pe	--	--	**1.00**	--	--	--	--	--	--	--
	/pɛ/	--	0.22	--	**0.77**	--	--	--	--	--	--
	/pæ/	--	--	--	--	**1.00**	--	--	--	--	--
	/pɒ/	--	--	--	--	--	**1.00**	--	--	--	--
	/pʌ/	--	--	--	--	--	0.02	**0.97**	--	0.01	--
	/po/	--	--	--	--	--	--	--	**1.00**	--	--
	/pʊ/	--	--	--	--	--	--	--	--	**1.00**	--
	/pu/	--	--	--	--	--	--	--	--	--	**1.00**

Table A3.3. English Model recognition of Voice 1 /g, kV/ stimuli (vowel duration included). Numbers reflect APPs of each token being a member of the corresponding category. Largest English APPs are in bold.

		APP scores of belonging to each English category									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended vowels produced in English by Stimulus Voice 1	/gi/	**1.00**	--	--	--	--	--	--	--	--	--
	/gɪ/	--	**1.00**	--	--	--	--	--	--	--	--
	/ge	0.08	--	**0.92**	--	--	--	--	--	--	--
	/gɛ/	--	0.40	--	**0.60**	--	--	--	--	--	--
	/gæ/	--	--	--	--	**1.00**	--	--	--	--	--
	/gɒ/	--	--	--	--	--	**1.00**	--	--	--	--
	/gʌ/	--	--	--	**0.60**	0.01	--	0.16	--	0.23	--
	/go/	--	--	--	--	--	--	--	**0.70**	--	0.30
	/gʊ/	--	0.09	--	0.01	--	--	--	--	**0.90**	--
	/gu/	--	--	--	--	--	--	--	--	--	**1.00**
	/ki/	**1.00**	--	--	--	--	--	--	--	--	--
	/kɪ/	--	**0.99**	--	0.01	--	--	--	--	--	--
	/ke	--	--	**1.00**	--	--	--	--	--	--	--
	/kɛ/	--	--	--	**0.97**	0.01	--	0.02	--	--	--
	/kæ/	--	--	--	--	**0.92**	0.02	0.06	--	--	--
	/kɒ/	--	--	--	--	--	**0.98**	0.02	--	--	--
	/kʌ/	--	--	--	--	0.01	0.01	**0.98**	--	--	--
	/ko/	--	--	--	--	--	--	--	**1.00**	--	--
	/kʊ/	--	--	--	0.08	--	--	0.18	--	**0.74**	--
	/ku/	--	--	--	--	--	--	--	--	--	**1.00**

Table A3.4. English Model recognition of Voice 1 /z, sV/ stimuli (vowel duration included). Numbers reflect APPs of each token being a member of the corresponding category. Largest English APPs are in bold.

Intended vowels produced in English by Stimulus Voice 1	APP scores of belonging to each English category									
	/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
/zi/	**1.00**	--	--	--	--	--	--	--	--	--
/zɪ/	--	**1.00**	--	--	--	--	--	--	--	--
/ze	--	--	**1.00**	--	--	--	--	--	--	--
/zɛ/	--	0.08	--	**0.84**	--	--	0.01	--	0.06	--
/zæ/	--	--	--	--	**0.98**	0.02	--	--	--	--
/zɒ/	--	--	--	--	--	**0.67**	0.12	--	0.21	--
/zʌ/	--	--	--	0.02	0.01	--	**0.91**	--	0.06	--
/zo/	--	--	--	--	--	--	--	**0.93**	--	0.07
/zʊ/	--	0.03	--	--	--	--	--	--	**0.97**	--
/zu/	--	--	--	--	--	--	--	--	--	**1.00**
/si/	**1.00**	--	--	--	--	--	--	--	--	--
/sɪ/	--	**1.00**	--	--	--	--	--	--	--	--
/se	--	--	**1.00**	--	--	--	--	--	--	--
/sɛ/	--	--	--	**0.57**	0.01	--	0.40	--	0.02	--
/sæ/	--	--	--	--	**0.98**	0.02	--	--	--	--
/sɒ/	--	--	--	--	--	**1.00**	--	--	--	--
/sʌ/	--	--	--	0.01	--	--	**0.86**	--	0.12	--
/so/	--	--	--	--	--	--	--	**1.00**	--	--
/sʊ/	--	--	--	--	--	--	--	--	**0.99**	--
/su/	--	--	--	--	--	--	--	--	--	**1.00**

Table A3.5. English Model recognition of Voice 1 /b, pV/ stimuli (vowel duration excluded). Numbers reflect APPs of each token being a member of the corresponding category. Largest English APPs are in bold.

		APP scores of belonging to each English category									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended vowels produced in English by Stimulus Voice 1	/bi/	**1.00**	--	--	--	--	--	--	--	--	--
	/bɪ/	--	**1.00**	--	--	--	--	--	--	--	--
	/be	--	--	**1.00**	--	--	--	--	--	--	--
	/bɛ/	--	0.16	--	**0.84**	--	--	--	--	--	--
	/bæ/	--	--	--	0.04	**0.82**	--	0.14	--	--	--
	/bɒ/	--	--	--	--	--	**0.90**	0.09	--	--	--
	/bʌ/	--	--	--	--	--	0.05	**0.73**	--	0.22	--
	/bo/	--	--	--	--	--	--	--	**1.00**	--	--
	/bʊ/	--	--	--	--	--	--	--	--	**1.00**	--
	/bu/	--	--	--	--	--	--	--	--	--	**1.00**
	/pi/	**1.00**	--	--	--	--	--	--	--	--	--
	/pɪ/	--	**0.98**	--	0.02	--	--	--	--	--	--
	/pe	--	--	**1.00**	--	--	--	--	--	--	--
	/pɛ/	--	--	--	**0.80**	0.11	--	0.09	--	--	--
	/pæ/	--	--	--	--	**0.88**	0.01	0.11	--	--	--
	/pɒ/	--	--	--	--	--	**0.91**	0.09	--	--	--
	/pʌ/	--	--	--	--	0.04	0.23	**0.72**	--	--	--
	/po/	--	--	--	--	--	--	--	**1.00**	--	--
	/pʊ/	--	--	--	--	--	--	--	--	**1.00**	--
	/pu/	--	--	--	--	--	--	--	--	--	**1.00**

Table A3.6. English Model recognition of Voice 2 /b, pV/ stimuli (vowel duration excluded). Numbers reflect APPs of each token being a member of the corresponding category. Largest English APPs are in bold.

		APP scores of belonging to each English category									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended vowels produced in English by Stimulus Voice 1	/bi/	**1.00**	--	--	--	--	--	--	--	--	--
	/bɪ/	--	**1.00**	--	--	--	--	--	--	--	--
	/be	--	--	**1.00**	--	--	--	--	--	--	--
	/bɛ/	--	0.06	--	**0.93**	--	--	--	--	--	--
	/bæ/	--	0.01	--	**0.85**	0.09	--	0.04	--	0.01	--
	/bɒ/	--	--	--	--	--	**0.44**	0.35	--	0.21	--
	/bʌ/	--	--	--	--	0.01	0.20	**0.74**	--	0.05	--
	/bo/	--	--	--	--	--	--	--	**1.00**	--	--
	/bʊ/	--	--	--	--	--	--	--	--	**1.00**	--
	/bu/	--	--	--	--	--	--	--	--	--	**1.00**
	/pi/	**1.00**	--	--	--	--	--	--	--	--	--
	/pɪ/	--	**0.98**	--	0.02	--	--	--	--	--	--
	/pe	--	--	**1.00**	--	--	--	--	--	--	--
	/pɛ/	--	0.20	--	**0.79**	--	--	--	--	0.01	--
	/pæ/	--	--	--	0.12	**0.87**	--	0.01	--	--	--
	/pɒ/	--	--	--	--	--	0.25	**0.57**	--	0.18	--
	/pʌ/	--	--	--	--	--	0.48	**0.52**	--	--	--
	/po/	--	--	--	--	--	--	--	**1.00**	--	--
	/pʊ/	--	--	--	--	--	--	--	--	**1.00**	--
	/pu/	--	--	--	--	--	--	--	--	--	**1.00**

Table A3.7. English Model recognition of Voice 1 /g, kV/ stimuli (vowel duration excluded). Numbers reflect APPs of each token being a member of the corresponding category. Largest English APPs are in bold.

Intended vowels produced in English by Stimulus Voice 1	APP scores of belonging to each English category									
	/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
/gi/	**1.00**	--	--	--	--	--	--	--	--	--
/gɪ/	--	**1.00**	--	--	--	--	--	--	--	--
/ge	0.31	0.06	**0.62**	--	--	--	--	--	--	--
/gɛ/	--	**0.60**	--	0.39	--	--	--	--	--	--
/gæ/	--	--	--	0.05	**0.84**	--	0.11	--	--	--
/gɒ/	--	--	--	--	--	**0.71**	0.28	--	0.01	--
/gʌ/	--	0.01	--	**0.59**	0.02	--	0.17	--	0.21	--
/go/	--	--	--	--	--	--	--	**0.71**	--	0.29
/gʊ/	--	0.15	--	0.01	--	--	--	--	**0.84**	--
/gu/	--	--	--	--	--	--	--	--	--	**1.00**
/ki/	**1.00**	--	--	--	--	--	--	--	--	--
/kɪ/	--	**0.99**	--	0.01	--	--	--	--	--	--
/ke	--	--	**1.00**	--	--	--	--	--	--	--
/kɛ/	--	0.01	--	**0.95**	0.02	--	0.02	--	--	--
/kæ/	--	--	--	--	**0.86**	0.02	0.12	--	--	--
/kɒ/	--	--	--	--	--	**0.94**	0.06	--	--	--
/kʌ/	--	--	--	--	0.08	0.05	**0.87**	--	--	--
/ko/	--	--	--	--	--	--	--	**1.00**	--	--
/kʊ/	--	--	--	0.07	--	--	0.18	--	**0.75**	--
/ku/	--	--	--	--	--	--	--	--	--	**1.00**

Table A3.8. English Model recognition of Voice 1 /z, sV/ stimuli (vowel duration excluded). Numbers reflect APPs of each token being a member of the corresponding category. Largest English APPs are in bold.

		APP scores of belonging to each English category									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended vowels produced in English by Stimulus Voice 1	/zi/	**1.00**	--	--	--	--	--	--	--	--	--
	/zɪ/	--	**1.00**	--	--	--	--	--	--	--	--
	/ze	--	--	**1.00**	--	--	--	--	--	--	--
	/zɛ/	--	0.17	--	**0.77**	--	--	0.01	--	0.05	--
	/zæ/	--	--	--	--	**0.75**	0.01	0.24	--	--	--
	/zɒ/	--	--	--	--	--	0.16	**0.43**	--	0.41	--
	/zʌ/	--	--	--	0.02	0.07	0.02	**0.85**	--	0.05	--
	/zo/	--	--	--	--	--	--	--	**0.96**	--	0.04
	/zʊ/	--	0.02	--	--	--	--	--	--	**0.97**	--
	/zu/	--	--	--	--	--	--	--	--	--	**1.00**
	/si/	**1.00**	--	--	--	--	--	--	--	--	--
	/sɪ/	--	**1.00**	--	--	--	--	--	--	--	--
	/se	--	--	**1.00**	--	--	--	--	--	--	--
	/sɛ/	--	--	--	**0.43**	0.26	--	0.29	--	0.01	--
	/sæ/	--	--	--	0.01	**0.77**	0.01	0.21	--	--	--
	/sɒ/	--	--	--	--	--	**0.54**	0.44	--	0.03	--
	/sʌ/	--	--	--	0.01	0.04	0.01	**0.82**	--	0.12	--
	/so/	--	--	--	--	--	--	--	**1.00**	--	--
	/sʊ/	--	--	--	--	--	--	--	--	**0.99**	--
	/su/	--	--	--	--	--	--	--	--	--	**1.00**

Table A3.9. Metamodel recognition of Voice 1 /b, pV/ stimuli (vowel duration excluded). Numbers reflect APPs of each token being a member of the corresponding category. Largest English APPs are in bold. Largest Mandarin competitor APPs are in italics.

	APP scores of belonging to each English and Mandarin category in the Metamodel																
	English										Mandarin						
Intended vowels produced in English by Stimulus Voice 1	/i/$_{e}$	/ɪ/$_{e}$	/e/$_{e}$	/ɛ/$_{e}$	/æ/$_{e}$	/ɒ/$_{e}$	/ʌ/$_{e}$	/o/$_{e}$	/ʊ/$_{e}$	/u/$_{e}$	/i/$_{m}$	/e/$_{m}$	/a/$_{m}$	/uə/$_{m}$	/o/$_{m}$	/ɤ/$_{m}$	/u/$_{m}$
/bi/	**0.36**	--	--	--	--	--	--	--	--	--	*0.64*	--	--	--	--	--	--
/bɪ/	--	**1.00**	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
/be	--	--	**0.68**	--	--	--	--	--	--	--	--	*0.32*	--	--	--	--	--
/bɛ/	--	0.08	--	**0.91**	--	--	--	--	0.01	--	--	--	--	--	--	--	--
/bæ/	--	--	--	0.02	**0.79**	--	0.09	--	--	--	--	--	*0.08*	--	--	--	--
/bɒ/	--	--	--	--	--	**0.17**	0.06	--	--	--	--	--	*0.77*	--	--	--	--
/bʌ/	--	--	--	--	0.01	0.03	**0.60**	--	0.23	--	--	--	*0.09*	--	--	0.03	--
/bo/	--	--	--	--	--	--	--	**0.59**	--	--	--	--	--	--	*0.41*	--	--
/bʊ/	--	--	--	--	--	--	--	--	**0.03**	--	--	--	--	0.04	--	*0.92*	--
/bu/	--	--	--	--	--	--	--	--	--	**1.00**	--	--	--	--	--	--	--
/pi/	**0.33**	--	--	--	--	--	--	--	--	--	*0.67*	--	--	--	--	--	--
/pɪ/	--	**0.96**	--	0.04	--	--	--	--	--	--	--	--	--	--	--	--	--
/pe	--	--	**0.65**	--	--	--	--	--	--	--	--	*0.35*	--	--	--	--	--
/pɛ/	--	--	--	**0.46**	0.17	--	0.35	--	0.01	--	--	--	--	--	--	--	--
/pæ/	--	--	--	--	**0.47**	0.04	0.12	--	--	--	--	--	*0.37*	--	--	--	--
/pɒ/	--	--	--	--	--	**0.11**	0.05	--	--	--	--	--	*0.84*	--	--	--	--
/pʌ/	--	--	--	--	0.01	0.13	**0.31**	--	--	--	--	--	*0.55*	--	--	--	--
/po/	--	--	--	--	--	--	--	**0.60**	--	--	--	--	--	--	*0.40*	--	--
/pʊ/	--	--	--	--	--	--	--	--	**0.22**	--	--	--	--	--	--	*0.78*	--
/pu/	--	--	--	--	--	--	--	--	--	**1.00**	--	--	--	--	--	--	--

Table A3.10. Metamodel recognition of Voice 2 /b, pV/ stimuli (vowel duration excluded). Numbers reflect APPs of each token being a member of the corresponding category. Largest English APPs are in bold. Largest Mandarin competitor APPs are in italics.

		APP scores of belonging to each English and Mandarin category in the Metamodel																
		English										Mandarin						
		/i/$_e$	/ɪ/$_e$	/e/$_e$	/ɛ/$_e$	/æ/$_e$	/ɒ/$_e$	/ʌ/$_e$	/ơ/$_e$	/ʊ/$_e$	/u/$_e$	/i/$_m$	/e/$_m$	/a/$_m$	/uə/$_m$	/o/$_m$	/ɤ/$_m$	/u/$_m$
Intended vowels produced in English by Stimulus Voice 1	/bi/	**0.83**	--	--	--	--	--	--	--	--	--	*0.17*	--	--	--	--	--	--
	/bɪ/	--	**1.00**	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	/be	--	--	**0.91**	--	--	--	--	--	--	--	--	*0.09*	--	--	--	--	--
	/bɛ/	--	0.05	--	**0.93**	0.01	--	0.01	--	0.01	--	--	--	--	--	--	--	--
	/bæ/	--	--	--	**0.77**	0.17	--	0.04	--	0.01	--	--	--	--	--	--	--	--
	/bɒ/	--	--	--	--	--	0.35	**0.43**	--	0.12	--	--	--	*0.09*	--	--	0.02	--
	/bʌ/	--	--	--	--	0.01	0.21	**0.67**	--	0.05	--	--	--	*0.05*	--	--	--	--
	/bo/	--	--	--	--	--	--	--	**0.92**	--	--	--	--	--	--	*0.08*	--	--
	/bʊ/	--	--	--	--	--	--	--	--	**0.35**	--	--	--	--	0.01	--	*0.63*	--
	/bu/	--	--	--	--	--	--	--	--	--	**1.00**	--	--	--	--	--	--	--
	/pi/	**0.92**	--	--	--	--	--	--	--	--	--	*0.08*	--	--	--	--	--	--
	/pɪ/	--	**0.96**	--	0.04	--	--	--	--	--	--	--	--	--	--	--	--	--
	/pe	--	--	**0.96**	--	--	--	--	--	--	--	--	*0.04*	--	--	--	--	--
	/pɛ/	--	0.14	--	**0.81**	--	--	0.01	--	0.04	--	--	--	--	--	--	--	--
	/pæ/	--	--	--	0.06	**0.91**	--	0.02	--	--	--	--	--	*0.01*	--	--	--	--
	/pɒ/	--	--	--	--	--	0.24	**0.53**	--	0.11	--	--	--	*0.11*	--	--	0.01	--
	/pʌ/	--	--	--	--	--	0.36	**0.33**	--	--	--	--	--	*0.30*	--	--	--	--
	/po/	--	--	--	--	--	--	--	**0.95**	--	--	--	--	--	--	*0.05*	--	--
	/pʊ/	--	--	--	--	--	--	--	--	**0.35**	--	--	--	--	0.01	--	*0.64*	--
	/pu/	--	--	--	--	--	--	--	--	--	**1.00**	--	--	--	--	--	--	--

Table A3.11. Metamodel recognition of Voice 1 /g, kV/ stimuli (vowel duration excluded). Numbers reflect APPs of each token being a member of the corresponding category. Largest English APPs are in bold. Largest Mandarin competitor APPs are in italics.

Intended vowels produced in English by Stimulus Voice 1	APP scores of belonging to each English and Mandarin category																
	English										Mandarin						
	/i/$_e$	/ɪ/$_e$	/e/$_e$	/ɛ/$_e$	/æ/$_e$	/ɒ/$_e$	/ʌ/$_e$	/o/$_e$	/ʊ/$_e$	/u/$_e$	/i/$_m$	/e/$_m$	/a/$_m$	/uə/$_m$	/o/$_m$	/ɤ/$_m$	/u/$_m$
/gi/	**0.38**	--	--	--	--	--	--	--	--	--	*0.62*	--	--	--	--	--	--
/gɪ/	--	**1.00**	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
/ge	0.20	0.06	**0.73**	--	--	--	--	--	--	--	*0.01*	--	--	--	--	--	--
/gɛ/	--	0.40	--	**0.60**	--	--	--	--	--	--	--	--	--	--	--	--	--
/gæ/	--	--	--	0.01	**0.84**	--	0.09	--	--	--	--	--	*0.05*	--	--	--	--
/gɒ/	--	--	--	--	--	**0.18**	0.14	--	--	--	--	--	*0.68*	--	--	--	--
/gʌ/	--	--	--	0.52	0.05	--	**0.26**	--	0.15	--	--	--	*0.01*	--	--	*0.01*	--
/go/	--	--	--	--	--	--	--	**0.95**	--	0.02	--	--	--	--	*0.03*	--	--
/gʊ/	--	0.11	--	0.01	--	--	--	--	**0.61**	--	--	--	--	--	--	*0.28*	--
/gu/	--	--	--	--	--	--	--	--	--	**1.00**	--	--	--	--	--	--	--
/ki/	**0.18**	--	--	--	--	--	--	--	--	--	*0.82*	--	--	--	--	--	--
/kɪ/	--	**0.99**	--	0.01	--	--	--	--	--	--	--	--	--	--	--	--	--
/ke	--	--	**0.58**	--	--	--	--	--	--	--	--	*0.42*	--	--	--	--	--
/kɛ/	--	--	--	**0.83**	0.04	--	0.10	--	0.02	--	--	--	--	--	--	--	--
/kæ/	--	--	--	--	**0.28**	0.04	0.08	--	--	--	--	--	*0.60*	--	--	--	--
/kɒ/	--	--	--	--	--	**0.22**	0.05	--	--	--	--	--	*0.73*	--	--	--	--
/kʌ/	--	--	--	--	0.04	0.08	**0.64**	--	--	--	--	--	*0.23*	--	--	--	--
/ko/	--	--	--	--	--	--	--	**0.49**	--	--	--	--	--	--	*0.51*	--	--
/kʊ/	--	--	--	0.04	--	--	0.11	--	**0.79**	--	--	--	--	--	--	*0.05*	--
/ku/	--	--	--	--	--	--	--	--	--	**1.00**	--	--	--	--	--	--	--

Table A3.12. Metamodel recognition of Voice 1 /z, sV/ stimuli (vowel duration excluded). Numbers reflect APPs of each token being a member of the corresponding category. Largest English APPs are in bold. Largest Mandarin competitor APPs are in italics.

		APP scores of belonging to each English and Mandarin category																
		English										Mandarin						
		/i/$_e$	/ɪ/$_e$	/e/$_e$	/ɛ/$_e$	/æ/$_e$	/ɒ/$_e$	/ʌ/$_e$	/o/$_e$	/ʊ/$_e$	/u/$_e$	/i/$_m$	/e/$_m$	/a/$_m$	/uə/$_m$	/o/$_m$	/ɤ/$_m$	/u/$_m$
Intended vowels produced in English by Stimulus Voice 1	/zi/	**0.76**	--	--	--	--	--	--	--	--	--	*0.23*	--	--	--	--	--	--
	/zɪ/	--	**1.00**	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
	/ze	--	--	**0.88**	--	--	--	--	--	--	--	--	*0.12*	--	--	--	--	--
	/zɛ/	--	0.19	--	**0.78**	0.01	--	--	--	0.02	--	--	--	--	--	--	--	--
	/zæ/	--	--	--	--	**0.48**	0.03	0.22	--	--	--	--	--	*0.26*	--	--	--	--
	/zɒ/	--	--	--	--	--	0.05	**0.29**	--	0.19	--	--	--	*0.35*	--	--	0.11	--
	/zʌ/	--	--	--	0.04	0.07	0.02	**0.73**	--	0.06	--	--	--	*0.09*	--	--	--	--
	/zo/	--	--	--	--	--	--	--	**0.96**	--	0.01	--	--	--	--	*0.03*	--	--
	/zʊ/	--	0.07	--	0.01	--	--	--	--	**0.50**	0.01	--	--	--	--	--	*0.41*	--
	/zu/	--	--	--	--	--	--	--	--	--	**1.00**	--	--	--	--	--	--	--
	/si/	**0.71**	--	--	--	--	--	--	--	--	--	*0.29*	--	--	--	--	--	--
	/sɪ/	--	**0.99**	--	--	--	--	--	--	--	0.01	--	--	--	--	--	--	--
	/se	--	--	**0.35**	--	--	--	--	--	--	--	--	*0.65*	--	--	--	--	--
	/sɛ/	--	--	--	**0.40**	0.22	--	0.31	--	0.02	--	--	--	*0.03*	--	--	--	--
	/sæ/	--	--	--	--	**0.54**	0.02	0.11	--	--	--	--	--	*0.33*	--	--	--	--
	/sɒ/	--	--	--	--	--	**0.09**	0.14	--	0.01	--	--	--	*0.76*	--	--	--	--
	/sʌ/	--	--	--	0.03	0.05	0.01	**0.72**	--	0.09	--	--	--	*0.08*	--	--	0.01	--
	/so/	--	--	--	--	--	--	--	**0.90**	--	--	--	--	--	--	*0.10*	--	--
	/sʊ/	--	--	--	--	--	--	--	--	**0.65**	--	--	--	--	--	--	*0.34*	--
	/su/	--	--	--	--	--	--	--	--	--	**1.00**	--	--	--	--	--	--	--

Appendix 4. Presentation order of English vowel categories in the training Demo Mode

Table A4. Presentation order of English vowel categories in Demo Mode. Within each step items were presented randomly.

Step	Vowel categories	Repetitions
1	/ɛ/ - /ɪ/	3
2	/ʊ/ - /ʌ/	3
3	/ɛ/ - /ɪ/- /ʊ/ - /ʌ/	3
4	/ɑ/ - /æ/	3
5	/ɛ/ - /ɪ/- /ʊ/ - /ʌ/- /ɑ/ - /æ/	3
6	/e/ - /i/	3
7	/o/ - /u/	3
8	/e/ - /i/ - /o/ - /u/	3
9	/ɑ/ - /æ/- /e/ - /i/ - /o/ - /u/	3
10	All categories	3

Appendix 5. Detailed statistics for the L2 English vowel training study

Table A5.1. ANOVA and Multivariate Tests comparing results on Generalization Test by Time, Consonant and Vowel, for Voice 1 only.

Source	Within-Subjects Measures	*df*	*F*	*p*
Time	Sphericity Assumed	1, 23	5.163	0.033
	Greenhouse-Geisser	1, 23	5.163	0.033
	Huynh-Feldt	1, 23	5.163	0.033
	Lower-bound	1, 23	5.163	0.033
Consonant	Sphericity Assumed	2, 46	26.322	0.000
	Greenhouse-Geisser	1.7356, 39.927	26.322	0.000
	Huynh-Feldt	2, 46	26.322	0.000
	Lower-bound	1, 23	26.322	0.000
Vowel	Sphericity Assumed	9, 207	18.665	0.000
	Greenhouse-Geisser	5.233, 120.351	18.665	0.000
	Huynh-Feldt	7.545, 173.528	18.665	0.000
	Lower-bound	1, 23	18.665	0.000
Consonant x Vowel	Sphericity Assumed	18, 414	4.502	0.000
	Greenhouse-Geisser	8.847, 203.475	4.502	0.000
	Huynh-Feldt	16.110, 370.539	4.502	0.000
	Lower-bound	1, 23	4.502	0.045

	Multivariate Measures			
Time	Wilks' Lambda	1, 23	5.163	0.033
Consonant	Wilks' Lambda	2, 22	24.891	0.000
Vowel	Wilks' Lambda	9, 15	23.494	0.000
Consonant x Vowel	Wilks' Lambda	18,6	2.486	0.132

Table A5.2. ANOVA and Multivariate Tests comparing results on Generalization Test by Time, Voice and Vowel, for /b, pV/ context only.

Source	Within-Subjects Measures	*df*	*F*	*p*
Time	Sphericity Assumed	1, 23	6.849	0.015
	Greenhouse-Geisser	1, 23	6.849	0.015
	Huynh-Feldt	1, 23	6.849	0.015
	Lower-bound	1, 23	6.849	0.015
Voice	Sphericity Assumed	1, 23	0.010	0.924
	Greenhouse-Geisser	1, 23	0.010	0.924
	Huynh-Feldt	1, 23	0.010	0.924
	Lower-bound	1, 23	0.010	0.924
Vowel	Sphericity Assumed	9, 207	19.711	0.000
	Greenhouse-Geisser	5.498, 126.457	19.711	0.000
	Huynh-Feldt	8.054, 185.231	19.711	0.000
	Lower-bound	1, 23	19.711	0.000
Vowel x Voice	Sphericity Assumed	9, 207	2.712	0.005
	Greenhouse-Geisser	5.139, 118.189	2.712	0.022
	Huynh-Feldt	7.368, 169.466	2.712	0.010
	Lower-bound	1,23	2.712	0.113
	Multivariate Measures			
Time	Wilks' Lambda	1, 23	6.849	0.015
Voice	Wilks' Lambda	1, 23	0.009	0.924
Vowel	Wilks' Lambda	9, 15	14.117	0.000
Vowel x Voice	Wilks' Lambda	9,15	2.858	0.035

Table A5.3. ANOVA results on Training Vowel Identification Tests (Lengthened or Natural) by Time and Vowel. Training groups are indicated separately.

Training Group	Source	Within-Subjects Measures	*df*	*F*	*p*
LVT	Time	Sphericity Assumed	1,10	16.251	0.002
		Greenhouse-Geisser	1,10	16.251	0.002
		Huynh-Feldt	1,10	16.251	0.002
		Lower-bound	1,10	16.251	0.002
	Vowel	Sphericity Assumed	9,90	22.472	0.000
		Greenhouse-Geisser	3.527,35.269	22.472	0.000
		Huynh-Feldt	5.684,56.844	22.472	0.000
		Lower-bound	1,10	22.472	0.001
SVT	Time	Sphericity Assumed	1,10	12.601	0.005
		Greenhouse-Geisser	1,10	12.601	0.005
		Huynh-Feldt	1,10	12.601	0.005
		Lower-bound	1,10	12.601	0.005
	Vowel	Sphericity Assumed	9,90	34.728	0.000
		Greenhouse-Geisser	4.430,44301	34.728	0.000
		Huynh-Feldt	8.390,83.901	34.728	0.000
		Lower-bound	1,10	34.728	0.000
DVT	Time	Sphericity Assumed	1,3	52.08333	0.005
		Greenhouse-Geisser	1,3	52.08333	0.005
		Huynh-Feldt	1,3	52.08333	0.005
		Lower-bound	1,3	52.08333	0.005
	Vowel	Sphericity Assumed	9,27	4.591356	0.001
		Greenhouse-Geisser	1.726,5.178	4.591356	0.074
		Huynh-Feldt	3.849,11.547	4.591356	0.0193
		Lower-bound	1,3	4.591356	0.122

Table A5.4. ANOVA and Multivariate Tests comparing results on Lengthened and Natural Identification Tests by Vowel Stimulus Length and Vowel.

Source	Within-Subjects Measures	*df*	*F*	*p*
Vowel Stimulus Length	Sphericity Assumed	1,23	0.064	0.803
	Greenhouse-Geisser	1,23	0.064	0.803
	Huynh-Feldt	1,23	0.064	0.803
	Lower-bound	1,23	0.064	0.803
Vowel	Sphericity Assumed	9,207	42.831	0.000
	Greenhouse-Geisser	5.215,119.593	42.831	0.000
	Huynh-Feldt	7.512,172.777	42.831	0.000
	Lower-bound	1,23	42.831	0.000
Vowel Stimulus Length x Vowel	Sphericity Assumed	9,207	4.560	0.000
	Greenhouse-Geisser	4.980114.543	4.560	0.001
	Huynh-Feldt	7.075,162.715	4.560	0.000
	Lower-bound	1,23	4.560	0.044
	Multivariate Measures			
Vowel Stimulus Length	Wilks' Lambda	1, 23	0.064	0.803
Vowel	Wilks' Lambda	9, 15	21.686	0.000
Vowel Stimulus Length x Vowel	Wilks' Lambda	9, 15	2.430	0.062

Table A5.5. ANOVA and Multivariate Tests comparing results on Generalization Identification Test at Time 2 with results on Generalization Delayed post-test by Consonant and Vowel for Voice 1 only.

Source	Within-Subjects Measures	*df*	*F*	*p*
Time	Sphericity Assumed	1,16	0.033	0.859
	Greenhouse-Geisser	1,16	0.033	0.859
	Huynh-Feldt	1,16	0.033	0.859
	Lower-bound	1,16	0.033	0.859
Consonant	Sphericity Assumed	2,32	22.079	0.000
	Greenhouse-Geisser	1.690,27.042	22.079	0.000
	Huynh-Feldt	1.986,31.780	22.079	0.000
	Lower-bound	1,16	22.079	0.000
Vowel	Sphericity Assumed	9,144	17.475	0.000
	Greenhouse-Geisser	5.136,82.176	17.475	0.000
	Huynh-Feldt	8.325,133.207	17.475	0.000
	Lower-bound	1,16	17.475	0.001
Consonant x Vowel	Sphericity Assumed	18,288	3.888	0.000
	Greenhouse-Geisser	6.463,103.406	3.888	0.001
	Huynh-Feldt	11.988,191.808	3.888	0.000
	Lower-bound	1,16	3.888	0.066

	Multivariate Measures			
Time	Wilks' Lambda	1, 16	0.033	0.859
Consonant	Wilks' Lambda	2, 15	18.663	0.000
Vowel	Wilks' Lambda	9, 8	9.306	0.002
Consonant x Vowel	Wilks' Lambda	too few	--	--

Table A5.6. ANOVA and Multivariate Tests comparing results on Generalization Test at Time 2 with results on Generalization Delayed post-test by Voice and Vowel.

Source	Within-Subjects Measures	*df*	*F*	*p*
Time	Sphericity Assumed	1,16	1.940	0.183
	Greenhouse-Geisser	1,16	1.940	0.183
	Huynh-Feldt	1,16	1.940	0.183
	Lower-bound	1,16	1.940	0.183
Voice	Sphericity Assumed	1,16	0.651	0.431
	Greenhouse-Geisser	1,16	0.651	0.431
	Huynh-Feldt	1,16	0.651	0.431
	Lower-bound	1,16	0.651	0.431
Vowel	Sphericity Assumed	9,144	15.652	0.000
	Greenhouse-Geisser	4.206,67.298	15.652	0.000
	Huynh-Feldt	6.250,99.998	15.652	0.000
	Lower-bound	1,16	15.652	0.001
Vowel x Voice	Sphericity Assumed	9,144	2.183	0.026
	Greenhouse-Geisser	4.163,66.611	2.183	0.078
	Huynh-Feldt	6.162,98.591	2.183	0.049
	Lower-bound	1,16	2.183	0.159
	Multivariate Measures			
Time	Wilks' Lambda	1, 16	1.940	0.183
Voice	Wilks' Lambda	1, 16	0.651	0.431
Vowel	Wilks' Lambda	9, 8	6.707	0.007
Vowel x Voice	Wilks' Lambda	9,15	1.497	0.290

Table A5.7. ANOVA results on Natural Vowel Identification on Test 2 and Delayed Post-test by Vowel.

Source	Within-Subjects Measures	*df*	*F*	*p*
Time	Sphericity Assumed	1,16	0.476	0.500
	Greenhouse-Geisser	1,16	0.476	0.500
	Huynh-Feldt	1,16	0.476	0.500
	Lower-bound	1,16	0.476	0.500
Vowel	Sphericity Assumed	9,144	45.588	0.000
	Greenhouse-Geisser	3.859,61.737	45.588	0.000
	Huynh-Feldt	5.556,88.891	45.588	0.000
	Lower-bound	1,16	45.588	0.000
Time x Vowel	Sphericity Assumed	9,144	2.991	0.003
	Greenhouse-Geisser	4.785, 76.568	2.991	0.017
	Huynh-Feldt	7.503,120.043	2.991	0.005
	Lower-bound	1,16	2.991	0.103
	Multivariate Measures			
Time	Wilks' Lambda	1, 16	.476	0.500
Vowel	Wilks' Lambda	9,8	18.141	0.000
Time x Vowel	Wilks' Lambda	9, 8	2.705	0.088

Table A5.8. ANOVA and Multivariate Tests comparing production results by Time, Consonant pair, and Vowel, in response to Voice 1 only, tested on the English Model with vowel duration excluded as a variable.

Source	Within-Subjects Measures	*df*	*F*	*p*
Time	Sphericity Assumed	1,20	1.847	0.189
	Greenhouse-Geisser	1,20	1.847	0.189
	Huynh-Feldt	1,20	1.847	0.189
	Lower-bound	1,20	1.847	0.189
Consonant	Sphericity Assumed	2,40	3.200	0.051
	Greenhouse-Geisser	1.631,32.626	3.200	0.063
	Huynh-Feldt	1.845,36.898	3.200	0.056
	Lower-bound	1,20	3.200	0.089
Vowel	Sphericity Assumed	9	21.569	0.000
	Greenhouse-Geisser	4.976,99.527	21.569	0.000
	Huynh-Feldt	7.154,143.080	21.569	0.000
	Lower-bound	1	21.569	0.000
Time * Consonant	Sphericity Assumed	2,40	4.291	0.021
	Greenhouse-Geisser	1.681,33.622	4.291	0.027
	Huynh-Feldt	1.910,38.194	4.291	0.022
	Lower-bound	1,20	4.291	0.051
Consonant * Vowel	Sphericity Assumed	18,360	9.167	0.000
	Greenhouse-Geisser	8.280,165.595	9.167	0.000
	Huynh-Feldt	15.371,307.424	9.167	0.000
	Lower-bound	1	9.167	0.007
	Multivariate Measures			
Time	Wilks' Lambda	1,20	1.847	0.189
Consonant	Wilks' Lambda	2, 19	2.886	0.080
Vowel	Wilks' Lambda	9,12	70.767	0.000
Time * Consonant	Wilks' Lambda	2, 19	7.214	0.005
Consonant * Vowel	Wilks' Lambda	18,3	10.833	0.037

Table A5.9. ANOVA and Multivariate Tests comparing production results for /b, pV/ only by Time, and Vowel, in response to Voice 1 only, tested on the English Model with vowel duration excluded as a variable.

Source	Within-Subjects Measures	*df*	*F*	*p*
Time	Sphericity Assumed	1,20	7.660	0.012
	Greenhouse-Geisser	1,20	7.660	0.012
	Huynh-Feldt	1,20	7.660	0.012
	Lower-bound	1,20	7.660	0.012
Vowel	Sphericity Assumed	9,180	28.682	0.000
	Greenhouse-Geisser	4.606,92.119	28.682	0.000
	Huynh-Feldt	6.453,129.051	28.682	0.000
	Lower-bound	1,20	28.682	0.000
Time * Vowel	Sphericity Assumed	9,180	2.272	0.020
	Greenhouse-Geisser	5.404,108,086	2.272	0.048
	Huynh-Feldt	8.009,160.176	2.272	0.025
	Lower-bound	1,20	2.272	0.147
	Multivariate Measures			
Time	Wilks' Lambda	1,20	7.660	0.012
Vowel	Wilks' Lambda	9,12	35.782	0.000
Time * Vowel	Wilks' Lambda	9,12	1.835	0.162

Table A5.10. ANOVA and Multivariate Tests comparing production results for /g, kV/ only by Time, and Vowel, in response to Voice 1 only, tested on the English Model with vowel duration excluded as a variable.

Source	Within-Subjects Measures	*df*	*F*	*p*
Time	Sphericity Assumed	1,20	0.112	0.741
	Greenhouse-Geisser	1,20	0.112	0.741
	Huynh-Feldt	1,20	0.112	0.741
	Lower-bound	1,20	0.112	0.741
Vowel	Sphericity Assumed	9,180	12.393	0.000
	Greenhouse-Geisser	5.635,112.708	12.393	0.000
	Huynh-Feldt	8.492,169.832	12.393	0.000
	Lower-bound	1,20	12.393	0.000
Time * Vowel	Sphericity Assumed	9,180	0.677	.730
	Greenhouse-Geisser	5.699,113.972	0.677	.730
	Huynh-Feldt	8.626,172.528	0.677	.730
	Lower-bound	1,20	0.677	.730
	Multivariate Measures			
Time	Wilks' Lambda	1,20	0.112	0.741
Vowel	Wilks' Lambda	9,12	48.123	0.000
Time * Vowel	Wilks' Lambda	9,12	0.564	0.802

Table A5.11. ANOVA and Multivariate Tests comparing production results for /z, sV/ only by Time, and Vowel, in response to Voice 1 only, tested on the English Model with vowel duration excluded as a variable.

Source	Within-Subjects Measures	*df*	*F*	*p*
Time	Sphericity Assumed	1,20	0.870	0.362
	Greenhouse-Geisser	1,20	0.870	0.362
	Huynh-Feldt	1,20	0.870	0.362
	Lower-bound	1,20	0.870	0.362
Vowel	Sphericity Assumed	9,180	12.283	0.000
	Greenhouse-Geisser	4.707,94.138	12.283	0.000
	Huynh-Feldt	6.640,132.808	12.283	0.000
	Lower-bound	1,20	12.283	0.000
Time * Vowel	Sphericity Assumed	9,180	1.781	0.074
	Greenhouse-Geisser	5.467,109.344	1.781	0.116
	Huynh-Feldt	8.139,167.775	1.781	0.083
	Lower-bound	1,20	1.781	0.197

	Multivariate Measures			
Time	Wilks' Lambda	1,20	0.870	0.362
Vowel	Wilks' Lambda	9,12	13.789	0.000
Time * Vowel	Wilks' Lambda	9,12	0.915	0.543

Table A5.12. ANOVA and Multivariate Tests comparing production results by Time, Voice, and Vowel, for the /b, pV/ context only, tested on the English Model with vowel duration excluded as a variable.

Source	Within-Subjects Measures	*df*	*F*	*p*
Time	Sphericity Assumed	1,20	9.463	.006
	Greenhouse-Geisser	1,20	9.463	.006
	Huynh-Feldt	1,20	9.463	.006
	Lower-bound	1,20	9.463	.006
Voice	Sphericity Assumed	1,20	0.202	0.658
	Greenhouse-Geisser	1,20	0.202	0.658
	Huynh-Feldt	1,20	0.202	0.658
	Lower-bound	1,20	0.202	0.658
Vowel	Sphericity Assumed	9,180	39.278	0.000
	Greenhouse-Geisser	4.247,84.935	39.278	0.000
	Huynh-Feldt	5.804,116.076	39.278	0.000
	Lower-bound	1,20	39.278	0.000
Voice * Vowel	Sphericity Assumed	9,180	4.515	0.000
	Greenhouse-Geisser	4.486,89.725	4.515	0.002
	Huynh-Feldt	6.233,124.660	4.515	0.000
	Lower-bound	1,20	4.515	0.046

	Multivariate Measures			
Time	Wilks' Lambda	1,20	9.463	0.006
Voice	Wilks' Lambda	1,20	0.202	0.658
Voice * TrainGrp	Wilks' Lambda	1,20	0.922	0.539
Vowel	Wilks' Lambda	9,12	78.359	0.000
Voice * Vowel	Wilks' Lambda	9,12	6.249	0.002

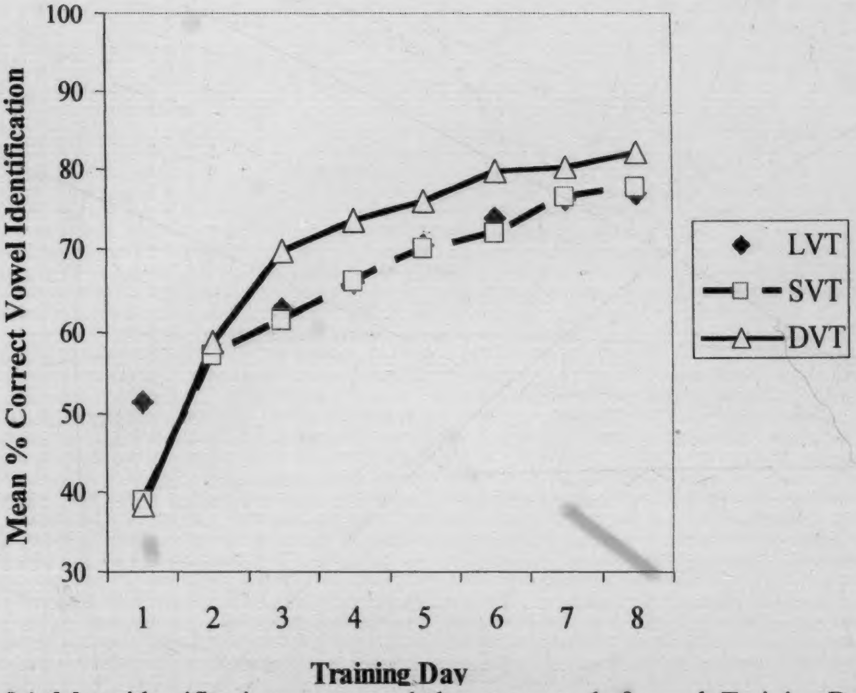


Figure A5.1. Mean identification scores pooled across vowels for each Training Day by each Training Group. Mean Pearson Correlation between groups, $r = .98$.

Appendix 6. Detailed statistics for the L2 English production results when vowel duration is included a variable in the CV English model

Table A6.1. Mean % correct vowel production recognition scores by CV context, stimulus Voice and Time. Vowel duration was included as a factor in the English CV pattern recognition model. Standard deviations are also provided.

	/b, pV/ Primary voice		/g, kV/ Primary voice		/z, sV/ Primary voice	
Group	LVT	SVT	LVT	SVT	LVT	SVT
Time 1	76.82	69.55	76.82	73.18	72.27	65.00
Time 2	80.00	75.68	72.50	69.77	72.50	66.59

	/b, pV/ (Secondary voice)		Average across groups and CVs
Group	LVT	SVT	
Time 1	78.40	77.30	73.67
Time 2	85.00	80.00	74.69

Table A6.2. ANOVA and Multivariate Tests comparing production results by Time, Consonant pair, and Vowel, in response to Voice 1 only, tested on the English Model with vowel duration included as a variable.

Source	Within-Subjects Measures	*df*	*F*	*p*
Time	Sphericity Assumed	1,20	0.122	0.730
	Greenhouse-Geisser	1,20	0.122	0.730
	Huynh-Feldt	1,20	0.122	0.730
	Lower-bound	1,20	0.122	0.730
Consonant	Sphericity Assumed	2,40	6.111	0.005
	Greenhouse-Geisser	1.622,32.447	6.111	0.009
	Huynh-Feldt	1.833,36.665	6.111	0.006
	Lower-bound	1,20	6.111	0.023
Vowel	Sphericity Assumed	9,180	15.480	0.000
	Greenhouse-Geisser	5.292,105.833	15.480	0.000
	Huynh-Feldt	7.779,155.580	15.480	0.000
	Lower-bound	1,20	15.480	0.001
Time * Consonant	Sphericity Assumed	2,40	7.479	0.002
	Greenhouse-Geisser	1.433,28.653	7.479	0.005
	Huynh-Feldt	1.590,31.796	7.479	0.004
	Lower-bound	1,20	7.479	0.013
Consonant * Vowel	Sphericity Assumed	18,360	9.188	0.000
	Greenhouse-Geisser	8.349,166.978	9.188	0.000
	Huynh-Feldt	15.593,311.862	9.188	0.000
	Lower-bound	1,20	9.188	0.007
	Multivariate Measures			
Time	Wilks' Lambda	1,20	0.122	0.730
Consonant	Wilks' Lambda	2,19	3.916	0.038
Vowel	Wilks' Lambda	9,12	37.665	0.000
Time * Consonant	Wilks' Lambda	2,19	14.392	0.000
Consonant * Vowel	Wilks' Lambda	18,3	9.478	0.044

Table A6.3. ANOVA and Multivariate Tests comparing production results for /b, pV/ only by Time, and Vowel, in response to Voice 1 only, tested on the English Model with vowel duration included as a variable.

Source	Within-Subjects Measures	*df*	*F*	*p*
Time	Sphericity Assumed	1,20	11.953	0.002
	Greenhouse-Geisser	1,20	11.953	0.002
	Huynh-Feldt	1,20	11.953	0.002
	Lower-bound	1,20	11.953	0.002
Vowel	Sphericity Assumed	9,180	25.990	0.000
	Greenhouse-Geisser	5.516,110.315	25.990	0.000
	Huynh-Feldt	8.240,164.796	25.990	0.000
	Lower-bound	1,20	25.990	0.000
Time * Vowel	Sphericity Assumed	9,180	3.932	0.000
	Greenhouse-Geisser	4.424,88.472	3.932	0.004
	Huynh-Feldt	6.119,122.390	3.932	0.001
	Lower-bound	1,20	3.932	0.061
	Multivariate Measures			
Time	Wilks' Lambda	1,20	11.953	0.002
Vowel	Wilks' Lambda	9,12	15.743	0.000
Time * Vowel	Wilks' Lambda	9,12	2.394	0.080

Table A6.4. ANOVA and Multivariate Tests comparing production results for /g, kV/ only by Time, and Vowel, in response to Voice 1 only, tested on the English Model with vowel duration included as a variable.

Source	Within-Subjects Measures	*df*	*F*	*p*
Time	Sphericity Assumed	1,20	2.746	0.113
	Greenhouse-Geisser	1,20	2.746	0.113
	Huynh-Feldt	1,20	2.746	0.113
	Lower-bound	1,20	2.746	0.113
Vowel	Sphericity Assumed	9,180	8.968	0.000
	Greenhouse-Geisser	6.001,120.017	8.968	0.000
	Huynh-Feldt	9.000,180.000	8.968	0.000
	Lower-bound	1,20	8.968	0.000
Time * Vowel	Sphericity Assumed	9,180	1.780	0.075
	Greenhouse-Geisser	5.403,108.067	1.780	0.004
	Huynh-Feldt	8.007,160.137	1.780	0.001
	Lower-bound	1,20	1.780	0.061

	Multivariate Measures			
Time	Wilks' Lambda	1,20	2.746	0.113
Vowel	Wilks' Lambda	9,12	15.662	0.000
Time * Vowel	Wilks' Lambda	9,12	0.849	0.589

Table A6.5. ANOVA and Multivariate Tests comparing production results for /z, sV/ only by Time, and Vowel, in response to Voice 1 only, tested on the English Model with vowel duration included as a variable.

Source	Within-Subjects Measures	*df*	*F*	*p*
Time	Sphericity Assumed	1,20	0.230	0.637
	Greenhouse-Geisser	1,20	0.230	0.637
	Huynh-Feldt	1,20	0.230	0.637
	Lower-bound	1,20	0.230	0.637
Vowel	Sphericity Assumed	9,180	9.484	0.000
	Greenhouse-Geisser	4.948,98.957	9.484	0.000
	Huynh-Feldt	7.099,141.977	9.484	0.000
	Lower-bound	1,20	9.484	0.000
Time * Vowel	Sphericity Assumed	9,180	0.748	0.075
	Greenhouse-Geisser	5.662,113.236	0.748	0.004
	Huynh-Feldt	8.548,170.956	0.748	0.001
	Lower-bound	1,20	0.748	0.061

	Multivariate Measures			
Time	Wilks' Lambda	1,20	0.230	0.637
Vowel	Wilks' Lambda	9,12	28.006	0.000
Time * Vowel	Wilks' Lambda	9,12	0.509	0.842

Table A6.6. ANOVA and Multivariate Tests comparing production results by Time, Voice, and Vowel, for the /b, pV/ context only, tested on the English Model with vowel duration included as a variable.

Source	Within-Subjects Measures	*df*	*F*	*p*
Time	Sphericity Assumed	1,20	5.568	0.029
	Greenhouse-Geisser	1,20	5.568	0.029
	Huynh-Feldt	1,20	5.568	0.029
	Lower-bound	1,20	5.568	0.029
Voice	Sphericity Assumed	1,20	11.523	0.003
	Greenhouse-Geisser	1,20	11.523	0.003
	Huynh-Feldt	1,20	11.523	0.003
	Lower-bound	1,20	11.523	0.003
Vowel	Sphericity Assumed	9,180	31.992	0.000
	Greenhouse-Geisser	5.062, 101.250	31.992	0.000
	Huynh-Feldt	7.322,146.443	31.992	0.000
	Lower-bound	1,20	31.992	0.000
Voice * Vowel	Sphericity Assumed	9,180	4.054	0.000
	Greenhouse-Geisser	4.725,94.501	4.054	0.003
	Huynh-Feldt	6.674,133.488	4.054	0.001
	Lower-bound	1,20	4.054	0.058

	Multivariate Measures			
Time	Wilks' Lambda	1,20	5.568	0.029
Voice	Wilks' Lambda	1,20	11.523	0.003
Voice * TrainGrp	Wilks' Lambda	1,20	5.803	0.026
Vowel	Wilks' Lambda	9,12	22.830	0.000
Voice * Vowel	Wilks' Lambda	9,12	2.371	0.082

Appendix 7. Identification confusion matrixes for Natural and Lengthened Vowel tests at Time 1 and Time 2 by each subset of participants who took each test at both times.

Table A7.1. Identification of English /b, pV/ stimuli on the Natural Vowel Test at Time 1 and Time 2, for the SVT and DVT groups only.

			Vowel identified by L2 learner								
			/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Time 1 (n=15)	Vowel presented as stimuli	/i/	**97.2**	1.2	0.2	--	0.8	0.2	0.2	0.2	0.2	--
		/ɪ/	4.2	**57.3**	10.7	18.0	1.0	0.2	2.5	0.5	5.7	--
		/e/	3.5	3.7	**83.7**	2.3	4.5	0.5	0.2	0.3	1.2	0.2
		/ɛ/	--	20.7	5.5	**40.5**	13.3	2.3	7.0	0.8	9.3	0.5
		/æ/	0.2	1.8	3.3	10.8	**32.3**	38.3	12.2	--	1.0	--
		/ɒ/	0.8	0.5	1.3	1.5	9.7	**69.0**	15.0	0.8	1.3	--
		/ʌ/	--	2.8	1.3	2.8	6.0	23.2	**46.0**	1.0	16.3	0.5
		/o/	0.2	1.8	0.3	0.7	0.8	0.2	1.2	**87.0**	4.3	3.5
		/ʊ/	0.5	0.8	0.3	1.7	0.3	0.7	7.5	6.3	**67.3**	14.5
		/u/	0.5	0.5	0.2	0.3	0.7	0.3	0.3	5.5	1.2	**90.5**
Total correct							67.1%					
Time 2 (n=15)	Vowel presented as stimuli	/i/	**98.3**	0.8	0.2	0.2	0.2	0.2	0.2	--	--	--
		/ɪ/	1.7	**66.8**	5.3	20.3	2.5	0.2	1.8	--	1.2	0.2
		/e/	0.5	1.3	**94.7**	1.5	1.5	0.2	--	0.3	--	--
		/ɛ/	0.2	15.0	1.7	**58.3**	15.2	0.8	5.5	--	3.2	0.2
		/æ/	0.2	0.2	0.2	11.7	**51.0**	27.0	8.7	0.5	0.7	--
		/ɒ/	--	0.2	0.5	0.2	8.5	**77.2**	12.5	0.3	0.7	--
		/ʌ/	0.3	0.5	--	2.3	5.3	19.7	**55.2**	0.3	16.0	0.3
		/o/	--	0.2	0.3	0.3	--	--	--	**94.7**	3.2	1.3
		/ʊ/	0.5	0.5	0.2	0.3	1.2	0.3	6.8	4.5	**78.3**	7.3
		/u/	0.5	--	--	0.3	0.2	0.3	0.2	4.7	1.3	**92.5**
Total correct							76.7%					

Table A7.2. Identification of English /b, pV/ stimuli on the Lengthened Vowel Test at Time 1 and Time 2, for the LVT group only.

			Vowel identified by L2 learner									
			/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Time 1 (n=11)	Vowel presented as stimuli	/i/	**92.7**	5.5	--	--	--	0.9	--	0.2	--	0.7
		/ɪ/	1.8	**57.3**	4.8	20.0	5.7	0.2	6.4	0.2	1.4	2.3
		/e/	0.7	3.0	**84.3**	5.0	2.0	--	0.7	0.5	--	3.9
		/ɛ/	--	27.3	7.7	**42.5**	11.8	0.9	5.5	0.7	3.4	0.2
		/æ/	--	3.0	4.3	8.9	**38.9**	29.3	14.3	0.7	0.5	0.2
		/ɒ/	--	0.2	--	0.5	6.8	**79.3**	12.5	0.5	--	0.2
		/ʌ/	--	2.3	1.1	5.0	8.4	29.8	**35.0**	0.5	17.7	0.2
		/o/	--	0.2	--	0.2	--	1.4	0.9	**92.3**	4.5	0.5
		/ʊ/	0.2	2.5	0.7	2.0	2.5	0.2	5.7	1.8	**75.9**	8.4
		/u/	0.5	0.2	--	0.7	0.5	2.5	1.1	2.7	5.2	**86.6**
Total correct							68.5%					
Time 2 (n=11)	Vowel presented as stimuli	/i/	**96.4**	3.4	0.2	--	--	--	--	--	--	--
		/ɪ/	2.0	**65.9**	5.7	18.2	5.5	0.2	2.0	--	0.5	--
		/e/	--	1.1	**93.6**	3.2	1.1	--	0.7	--	0.2	--
		/ɛ/	--	20.0	3.2	**53.2**	18.2	0.9	3.4	--	1.1	--
		/æ/	0.2	1.4	0.7	7.0	**58.6**	22.0	10.0	--	--	--
		/ɒ/	--	0.2	--	0.5	8.4	**83.2**	6.6	0.9	0.2	--
		/ʌ/	--	1.1	0.2	1.8	6.4	26.1	**47.7**	0.2	16.4	--
		/o/	0.5	--	0.2	--	0.5	0.2	0.5	**96.8**	1.4	--
		/ʊ/	0.2	0.9	--	0.9	1.6	--	8.2	1.4	**78.9**	8.0
		/u/	1.4	0.2	--	--	0.7	--	0.2	1.8	5.0	**90.7**
Total correct							76.5%					

Appendix 8. Detailed statistics for the naturalistic L2 English vowel production data

Table A8.1. ANOVA and Multivariate test results for naturalistic L2 English vowel production data including vowel duration as a variable.

Source	Within-Subjects Measures	*df*	*F*	*p*
Time	Sphericity Assumed	5,95	3.976	0.003
	Greenhouse-Geisser	3.682,69.966	3.976	0.007
	Huynh-Feldt	4.678,88.874	3.976	0.003
	Lower-bound	1,19	3.976	0.061
Vowel	Sphericity Assumed	9,171	15.603	0.000
	Greenhouse-Geisser	5.989,113.791	15.603	0.000
	Huynh-Feldt	9,171	15.603	0.000
	Lower-bound	1,19	15.603	0.001
	Multivariate Measures			
Time	Wilks' Lambda	5,15	3.666	.023
Vowel	Wilks' Lambda	9,11	6.780	.002

Table A8.2. Naturalistic group's Time 1 L2 English CVC productions, tested on the CVC English Model with vowel duration included as a variable. Values represent percentages of intended vowels recognized as belonging to each English vowel category.

		Vowel recognized by CVC English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended English vowels repeated in response to auditory stimuli	/i/	**75.0**	2.5	22.5	--	--	--	--	--	--	--
	/ɪ/	--	**25.0**	12.5	57.5	5.0	--	--	--	--	--
	/e/	2.5	25.0	**67.5**	5.0	--	--	--	--	--	--
	/ɛ/	--	2.5	--	**65.0**	27.5	2.5	--	--	2.5	--
	/æ/	--	--	--	12.5	**75.0**	2.5	10.0	--	--	--
	/ɒ/	--	--	--	--	25.0	**70.0**	2.5	2.5	--	--
	/ʌ/	--	--	--	--	20.0	10.0	**67.5**	--	2.5	--
	/o/	--	--	--	--	--	--	--	**92.5**	7.5	--
	/ʊ/	--	--	--	--	--	10.0	2.5	40.0	**47.5**	--
	/u/	--	--	--	--	--	--	--	62.5	7.5	**30.0**
Total correct						61.5%					

Table A8.3. Naturalistic group's Time 2 L2 English CVC productions, tested on the CVC English Model with vowel duration included as a variable. Values represent percentages of intended vowels recognized as belonging to each English vowel category.

		Vowel recognized by CVC English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended English vowels repeated in response to auditory stimuli	/i/	**72.5**	--	27.5	--	--	--	--	--	--	--
	/ɪ/	2.5	**25.0**	15.0	50.0	5.0	--	2.5	--	--	--
	/e/	--	17.5	**72.5**	5.0	2.5	--	2.5	--	--	--
	/ɛ/	--	2.5	--	**65.0**	25.0	--	7.5	--	--	--
	/æ/	--	--	--	15.0	**72.5**	2.5	10.0	--	--	--
	/ɒ/	--	--	--	--	25.0	**75.0**	--	--	--	--
	/ʌ/	--	--	--	12.5	32.5	17.5	**37.5**	--	--	--
	/o/	--	--	--	--	--	2.5	--	**87.5**	2.5	7.5
	/ʊ/	--	--	--	--	--	10.0	--	45.0	**40.0**	5.0
	/u/	--	--	--	--	--	--	--	47.5	2.5	**50.0**
Total correct						59.8%					

Table A8.4. Naturalistic group's Time 3 L2 English CVC productions, tested on the CVC English Model with vowel duration included as a variable. Values represent percentages of intended vowels recognized as belonging to each English vowel category.

		Vowel recognized by CVC English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended English vowels repeated in response to auditory stimuli	/i/	**92.5**	--	7.5	--	--	--	--	--	--	--
	/ɪ/	--	**27.5**	7.5	60.0	5.0	--	--	--	--	--
	/e/	5.0	10.0	**82.5**	2.5	--	--	--	--	--	--
	/ɛ/	--	--	--	**67.5**	30.0	--	2.5	--	--	--
	/æ/	--	--	--	7.5	**75.0**	10.0	7.5	--	--	--
	/ɒ/	--	--	--	--	20.0	**77.5**	--	2.5	--	--
	/ʌ/	--	--	--	2.5	27.5	17.5	**52.5**	--	--	--
	/o/	--	--	--	--	--	--	--	**95.0**	--	5.0
	/ʊ/	--	--	--	--	--	5.0	2.5	47.5	**42.5**	2.5
	/u/	--	--	--	--	--	--	--	57.5	5.0	**37.5**
Total correct						65.0%					

Table A8.5. Naturalistic group's Time 4 L2 English CVC productions, tested on the CVC English Model with vowel duration included as a variable. Values represent percentages of intended vowels recognized as belonging to each English vowel category.

		Vowel recognized by CVC English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended English vowels repeated in response to auditory stimuli	/i/	**77.5**	--	22.5	--	--	--	--	--	--	--
	/ɪ/	5.0	**35.0**	10.0	37.5	7.5	--	5.0	--	--	--
	/e/	2.5	7.5	**90.0**	--	--	--	--	--	--	--
	/ɛ/	--	--	--	**60.0**	40.0	--	--	--	--	--
	/æ/	--	--	--	2.5	**87.5**	5.0	5.0	--	--	--
	/ɒ/	--	--	--	--	7.5	**87.5**	2.5	2.5	--	--
	/ʌ/	--	--	--	2.5	35.0	17.5	**35.0**	--	10.0	--
	/o/	--	--	--	--	--	2.5	--	**92.5**	--	5.0
	/ʊ/	--	--	--	--	2.5	--	2.5	52.5	**37.5**	5.0
	/u/	--	--	--	--	--	--	--	57.5	5.0	**37.5**
Total correct						64.0%					

Table A8.6. Naturalistic group's Time 5 L2 English CVC productions, tested on the CVC English Model with vowel duration included as a variable. Values represent percentages of intended vowels recognized as belonging to each English vowel category.

		Vowel recognized by CVC English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended English vowels repeated in response to auditory stimuli	/i/	**87.5**	--	12.5	--	--	--	--	--	--	--
	/ɪ/	--	**37.5**	12.5	45.0	5.0	--	--	--	--	--
	/e/	--	5.0	**92.5**	2.5	--	--	--	--	--	--
	/ɛ/	--	2.5	--	**60.0**	35.0	--	2.5	--	--	--
	/æ/	--	--	--	7.5	**85.0**	7.5	--	--	--	--
	/ɒ/	--	--	--	--	12.5	**85.0**	--	2.5	--	--
	/ʌ/	--	--	--	2.5	42.5	2.5	**52.5**	--	--	--
	/o/	--	--	--	--	--	2.5	--	**95.0**	--	2.5
	/ʊ/	--	--	--	--	--	2.5	--	37.5	**57.5**	2.5
	/u/	--	--	--	--	--	--	--	50.0	7.5	**42.5**
Total correct						69.5%					

Table A8.7. Naturalistic group's Time 6 L2 English CVC productions, tested on the CVC English Model with vowel duration included as a variable. Values represent percentages of intended vowels recognized as belonging to each English vowel category.

		Vowel recognized by CVC English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended English vowels repeated in response to auditory stimuli	/i/	**90.0**	--	10.0	--	--	--	--	--	--	--
	/ɪ/	2.5	**42.5**	20.0	35.0	--	--	--	--	--	--
	/e/	2.5	2.5	**92.5**	2.5	--	--	--	--	--	--
	/ɛ/	--	--	--	**55.0**	45.0	--	--	--	--	--
	/æ/	--	--	--	10.0	**87.5**	2.5	--	--	--	--
	/ɒ/	--	--	--	--	5.0	**90.0**	--	5.0	--	--
	/ʌ/	--	--	--	2.5	30.0	12.5	**50.0**	--	5.0	--
	/o/	--	--	--	--	--	2.5	--	**92.5**	2.5	2.5
	/ʊ/	--	2.5	--	--	--	12.5	--	45.0	**35.0**	5.0
	/u/	--	--	--	--	--	--	--	35.0	7.5	**57.5**
Total correct						69.3%					

Table A8.8. Training group's Time 1 L2 English CV productions in response to Voice 1 /b, pV/ stimuli, tested on the CV English Model with vowel duration included as a variable. Values represent percentages of intended vowels recognized as belonging to each English vowel category.

		Vowel recognized by CV English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended English vowels repeated in response to auditory stimuli	/i/	**97.7**	--	2.3	--	--	--	--	--	--	--
	/ɪ/	1.1	**51.1**	10.2	36.4	--	--	1.1	--	--	--
	/e/	1.1	--	**96.6**	1.1	1.1	--	--	--	--	--
	/ɛ/	--	6.8	--	**73.9**	12.5	1.1	5.7	--	--	--
	/æ/	--	--	--	11.4	**68.2**	13.6	6.8	--	--	--
	/ɒ/	--	--	--	--	3.4	**88.6**	4.5	2.3	1.1	--
	/ʌ/	--	--	--	1.1	3.4	35.2	**58.0**	1.1	1.1	--
	/o/	--	--	--	--	--	3.4	--	**90.9**	5.7	--
	/ʊ/	--	--	--	--	--	10.2	4.5	8.0	**76.1**	1.1
	/u/	--	--	1.1	--	--	--	--	22.7	45.5	**30.7**
Total correct						73.2%					

Table A8.9. Training group's Time 2 L2 English CV productions in response to Voice 1 /b, pV/ stimuli, tested on the CV English Model with vowel duration included as a variable. Values represent percentages of intended vowels recognized as belonging to each English vowel category.

		Vowel recognized by CV English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended English vowels repeated in response to auditory stimuli	/i/	**96.6**	--	3.4	--	--	--	--	--	--	--
	/ɪ/	--	**71.6**	4.5	22.7	1.1	--	--	--	--	--
	/e/	4.5	--	**95.5**	--	--	--	--	--	--	--
	/ɛ/	--	9.1	1.1	**75.0**	11.4	--	3.4	--	--	--
	/æ/	--	--	--	6.8	**78.4**	13.6	--	--	1.1	--
	/ɒ/	--	--	--	--	2.3	**96.6**	--	1.1	--	--
	/ʌ/	--	--	--	--	2.3	28.4	**63.6**	--	5.7	--
	/o/	--	--	--	--	--	3.4	--	**90.9**	4.5	1.1
	/ʊ/	--	--	--	--	--	9.1	10.2	--	**80.7**	--
	/u/	--	--	--	--	--	--	--	27.3	43.2	**29.5**
Total correct						77.8%					

Appendix 9. Alternate statistics for the naturalistic L2 English vowel data and comparison with the training study data when vowel duration is excluded as a variable

Table A9.1. Recognition of English production tokens by vowel tested against the CVC English Model trained and tested on native speaker English productions with vowel duration excluded as a variable. Values represent percentages of intended vowels recognized as belonging to each English vowel category.

		Vowel recognized by CVC English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended	/i/	**95.5**	--	4.5	--	--	--	--	--	--	--
English	/ɪ/	--	**95.3**	--	4.7	--	--	--	--	--	--
vowels	/e/	6.1	--	**93.9**	--	--	--	--	--	--	--
repeated in	/ɛ/	--	3.1	--	**89.1**	7.8	--	--	--	--	--
response to	/æ/	--	--	--	7.8	**92.2**	--	--	--	--	--
auditory	/ɒ/	--	--	--	--	--	**98.5**	1.5	--	--	--
stimuli	/ʌ/	--	--	--	--	3.2	1.6	**90.3**	--	4.8	--
	/o/	--	--	--	--	--	--	--	**98.5**	1.5	--
	/ʊ/	--	--	--	--	--	4.7	3.1	1.6	**89.1**	1.6
	/u/	--	1.5	--	--	--	--	--	--	3.0	**95.5**
Total correct		93.9% (94.9% with vowel duration cue)									

Table A9.2. ANOVA and Multivariate test results for naturalistic L2 English vowel production data excluding vowel duration as a variable.

Source	Within-Subjects Measures	*df*	*F*	*p*
Time	Sphericity Assumed	5,95	2.915	0.017
	Greenhouse-Geisser	3.642,69.194	2.915	0.031
	Huynh-Feldt	4.612,87.663	2.915	0.020
	Lower-bound	1,19	2.915	0.104
Vowel	Sphericity Assumed	9,171	11.207	0.000
	Greenhouse-Geisser	6.030,114.575	11.207	0.000
	Huynh-Feldt	9,171	11.207	0.000
	Lower-bound	1,19	11.207	0.003
	Multivariate Measures			
Time	Wilks' Lambda	5,15	1.777	0.178
Vowel	Wilks' Lambda	9,11	6.679	0.002

Table A9.3. Mean percent correctly recognized vowel productions over time for the naturalistic group's L2 productions (top panel), in contrast to the trained group's L2 productions described in Chapters 4 and 5 (bottom panel). Vowel duration was excluded as a variable in both the CVC English and CV English Models.

	Naturalistic learners' /b,pVt/ production data (n= 20)	
Time	% Correct *(SD)*	% Improvement from previous time *(SD)*
1	59.5 *(11.11)*	
2	57.75 *(13.91)*	-1.75 *(12.90)*
3	62.25 *(10.44)*	4.50 *(12.96)*
4	62.25 *(10.44)*	0.00 *(11.00)*
5	64.75 *(13.62)*	2.50 *(12.72)*
6	68.00 *(10.44)*	3.25 *(16.24)*

	Trained learners' /b, pV/ production data (n=22) in response to Voice 1	
Time	% Correct *(SD)*	% Improvement from previous time *(SD)*
1	67.61 *(9.11)*	
2	73.52 *(10.63)*	5.91 *(9.79)*

Table A9.4. Summary of mean percent correctly recognized L2 English productions for each English vowel, contrasting naturalistic vowel learning study (from Times 1-6) with trained vowel learning results (from Times 1-2). Vowel duration excluded as a variable.

	Naturalistic Vowel Learning Study (n=20)							Vowel Training Study (n=22)		
Vowel	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6	Difference from Time 1 to 6 (approx. 10 months)	Time 1	Time 2	Difference from Time 1 to 2 (approx. 3 weeks)
/i/	85	77.5	90	80	90	90	5	97.7	98.9	1.2
/ɪ/	25	37.5	35	40	40	45	20	52.3	71.6	19.3
/e/	67.5	65	80	77.5	85	85	17.5	94.3	93.2	-1.1
/ɛ/	75	67.5	75	72.5	65	72.5	-2.5	76.1	70.5	-5.6
/æ/	50	60	50	65	60	70	20	40.9	56.8	15.9
/ɒ/	60	67.5	70	77.5	82.5	85	25	80.7	92.0	11.3
/ʌ/	42.5	30	45	27.5	35	42.5	0	29.5	42.0	12.5
/o/	90	80	92.5	95	87.5	90	0	92	88.6	-3.4
/ʊ/	60	45	47.5	45	62.5	45	-15	81.8	86.4	4.6
/u/	40	47.5	37.5	42.5	40	55	15	30.7	35.2	4.5
Mean	59.5	57.75	62.25	62.25	64.75	68	8.5	67.6	73.5	5.9

Table A9.5. Naturalistic group's Time 1 L2 English CVC productions, tested on the CVC English Model with vowel duration excluded as a variable. Values represent percentages of intended vowels recognized as belonging to each English vowel category.

		Vowel recognized by CVC English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended English vowels repeated in response to auditory stimuli	/i/	**85.0**	--	15.0	--	--	--	--	--	--	--
	/ɪ/	--	**25.0**	12.5	60.0	2.5	--	--	--	--	--
	/e/	--	30.0	**67.5**	2.5	--	--	--	--	--	--
	/ɛ/	--	2.5	--	**75.0**	15.0	2.5	2.5	--	2.5	--
	/æ/	--	--	--	32.5	**50.0**	2.5	15.0	--	--	--
	/ɒ/	--	--	--	--	22.5	**60.0**	12.5	2.5	2.5	--
	/ʌ/	--	--	--	--	40.0	15.0	**42.5**	--	2.5	--
	/o/	--	--	--	--	--	--	--	**90.0**	10.0	--
	/ʊ/	--	--	--	--	--	--	2.5	37.5	**60.0**	--
	/u/	--	--	--	--	--	--	--	50.0	10.0	**40.0**
Total correct						59.5%					

Table A9.6. Naturalistic group's Time 2 L2 English CVC productions, tested on the CVC English Model with vowel duration excluded as a variable. Values represent percentages of intended vowels recognized as belonging to each English vowel category.

		Vowel recognized by CVC English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended English vowels repeated in response to auditory stimuli	/i/	**77.5**	--	22.5	--	--	--	--	--	--	--
	/ɪ/	--	**37.5**	12.5	47.5	2.5	--	--	--	--	--
	/e/	2.5	25.0	**65.0**	5.0	--	--	2.5	--	--	--
	/ɛ/	--	2.5	--	**67.5**	22.5	--	7.5	--	--	--
	/æ/	--	--	--	22.5	**60.0**	--	17.5	--	--	--
	/ɒ/	--	--	--	--	17.5	**67.5**	7.5	2.5	5.0	--
	/ʌ/	--	--	--	7.5	42.5	20.0	**30.0**	--	--	--
	/o/	--	--	--	--	--	2.5	--	**80.0**	10.0	7.5
	/ʊ/	--	--	--	--	--	7.5	--	42.5	**45.0**	5.0
	/u/	--	--	--	--	--	--	--	47.5	5.0	**47.5**
Total correct						57.8%					

Table A9.7. Naturalistic group's Time 3 L2 English CVC productions, tested on the CVC English Model with vowel duration excluded as a variable. Values represent percentages of intended vowels recognized as belonging to each English vowel category.

		Vowel recognized by CVC English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended English vowels repeated in response to auditory stimuli	/i/	**90.0**	--	10.0	--	--	--	--	--	--	--
	/ɪ/	--	**35.0**	5.0	57.5	2.5	--	--	--	--	--
	/e/	5.0	15.0	**80.0**	--	--	--	--	--	--	--
	/ɛ/	--	--	--	**75.0**	22.5	--	2.5	--	--	--
	/æ/	--	--	--	25.0	**50.0**	5.0	20.0	--	--	--
	/ɒ/	--	--	--	--	17.5	**70.0**	2.5	2.5	7.5	--
	/ʌ/	--	--	--	2.5	32.5	20.0	**45.0**	--	--	--
	/o/	--	--	--	--	--	--	--	**92.5**	2.5	5.0
	/ʊ/	--	--	--	--	--	2.5	2.5	42.5	**47.5**	5.0
	/u/	--	--	--	--	--	--	--	57.5	5.0	**37.5**
Total correct						62.3%					

Table A9.8. Naturalistic group's Time 4 L2 English CVC productions, tested on the CVC English Model with vowel duration excluded as a variable. Values represent percentages of intended vowels recognized as belonging to each English vowel category.

		Vowel recognized by CVC English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended English vowels repeated in response to auditory stimuli	/i/	**80.0**	--	20.0	--	--	--	--	--	--	--
	/ɪ/	2.5	**40.0**	10.0	40.0	2.5	--	5.0	--	--	--
	/e/	7.5	15.0	**77.5**	--	--	--	--	--	--	--
	/ɛ/	--	--	--	**72.5**	27.5	--	--	--	--	--
	/æ/	--	--	--	22.5	**65.0**	2.5	10.0	--	--	--
	/ɒ/	--	--	--	--	10.0	**77.5**	7.5	2.5	2.5	--
	/ʌ/	--	--	--	--	42.5	20.0	**27.5**	--	10.0	--
	/o/	--	--	--	--	--	2.5	--	**95.0**	--	2.5
	/ʊ/	--	--	--	--	--	--	5.0	45.0	**45.0**	5.0
	/u/	--	--	--	--	--	--	--	52.5	5.0	**42.5**
Total correct						62.3%					

Table A9.9. Naturalistic group's Time 5 L2 English CVC productions, tested on the CVC English Model with vowel duration excluded as a variable. Values represent percentages of intended vowels recognized as belonging to each English vowel category.

		Vowel recognized by CVC English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended English vowels repeated in response to auditory stimuli	/i/	**90.0**	--	10.0	--	--	--	--	--	--	--
	/ɪ/	--	**40.0**	10.0	50.0	--	--	--	--	--	--
	/e/	2.5	12.5	**85.0**	--	--	--	--	--	--	--
	/ɛ/	--	2.5	--	**65.0**	25.0	--	7.5	--	--	--
	/æ/	--	--	--	22.5	**60.0**	7.5	10.0	--	--	--
	/ɒ/	--	--	--	--	10.0	**82.5**	5.0	2.5	--	--
	/ʌ/	--	--	--	--	50.0	15.0	**35.0**	--	--	--
	/o/	--	--	--	--	--	5.0	--	**87.5**	2.5	5.0
	/ʊ/	--	--	--	--	--	2.5	--	32.5	**62.5**	2.5
	/u/	--	--	--	--	--	--	--	50.0	10.0	**40.0**
Total correct						64.8%					

Table A9.10. Naturalistic group's Time 6 L2 English CVC productions, tested on the CVC English Model with vowel duration excluded as a variable. Values represent percentages of intended vowels recognized as belonging to each English vowel category.

		Vowel recognized by CVC English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended English vowels repeated in response to auditory stimuli	/i/	**90.0**	--	10.0	--	--	--	--	--	--	--
	/ɪ/	--	**45.0**	22.5	30.0	2.5	--	--	--	--	--
	/e/	2.5	12.5	**85.0**	--	--	--	--	--	--	--
	/ɛ/	--	--	--	**72.5**	27.5	--	--	--	--	--
	/æ/	--	--	--	27.5	**70.0**	2.5	--	--	--	--
	/ɒ/	--	--	--	--	2.5	**85.0**	7.5	2.5	2.5	--
	/ʌ/	--	--	--	--	40.0	15.0	**42.5**	--	2.5	--
	/o/	--	--	--	--	--	2.5	--	**90.0**	2.5	5.0
	/ʊ/	--	--	--	--	--	10.0	--	40.0	**45.0**	5.0
	/u/	--	--	--	--	--	--	--	35.0	10.0	**55.0**
Total correct						68%					

Table A9.11. Training group's Time 1 L2 English CV productions in response to Voice 1 /b, pV/ stimuli, tested on the CV English Model with vowel duration excluded as a variable. Values represent percentages of intended vowels recognized as belonging to each English vowel category.

		Vowel recognized by CV English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended English vowels repeated in response to auditory stimuli	/i/	**97.7**	--	2.3	--	--	--	--	--	--	--
	/ɪ/	--	**52.3**	13.6	33.0	--	--	1.1	--	--	--
	/e/	3.4	1.1	**94.3**	1.1	--	--	--	--	--	--
	/ɛ/	--	6.8	--	**76.1**	12.5	2.3	2.3	--	--	--
	/æ/	--	1.1	--	33.0	**40.9**	11.4	13.6	--	--	--
	/ɒ/	--	--	--	--	3.4	**80.7**	11.4	2.3	2.3	--
	/ʌ/	--	--	--	1.1	6.8	60.2	**29.5**	1.1	1.1	--
	/o/	--	--	--	--	--	--	--	**92.0**	8.0	--
	/ʊ/	--	--	--	--	--	4.5	4.5	9.1	**81.8**	--
	/u/	--	1.1	--	--	--	--	--	19.3	48.9	**30.7**
Total correct						67.6%					

Table A9.12. Training group's Time 2 L2 English CV productions in response to Voice 1 /b, pV/ stimuli, tested on the CV English Model with vowel duration excluded as a variable. Values represent percentages of intended vowels recognized as belonging to each English vowel category.

		Vowel recognized by CV English pattern recognition model									
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɒ/	/ʌ/	/o/	/ʊ/	/u/
Intended English vowels repeated in response to auditory stimuli	/i/	**98.9**	--	1.1	--	--	--	--	--	--	--
	/ɪ/	--	**71.6**	5.7	21.6	1.1	--	--	--	--	--
	/e/	6.8	--	**93.2**	--	--	--	--	--	--	--
	/ɛ/	--	11.4	1.1	**70.5**	14.8	--	2.3	--	--	--
	/æ/	--	--	--	19.3	**56.8**	12.5	10.2	--	1.1	--
	/ɒ/	--	--	--	--	2.3	**92.0**	2.3	--	3.4	--
	/ʌ/	--	--	--	--	5.7	47.7	**42.0**	--	4.5	--
	/o/	--	--	--	--	--	--	--	**88.6**	10.2	1.1
	/ʊ/	--	--	--	--	--	6.8	6.8	--	**86.4**	--
	/u/	--	--	--	--	--	--	--	19.3	45.5	**35.2**
Total correct						73.5%					